S

Scaling Limits of Large Systems of Non-linear Partial Differential Equations

D. Benedetto, M. Pulvirenti
Dipartimento di Matematica, Università di Roma
'La Sapienza', Roma, Italy

Article Outline

Glossary

Scaling limits A scaling limit denotes a procedure to reduce the degree of complexity of a large particle system. It consists in scaling the space-time variables and, possibly, other quantities (like the interaction potential or the density), in order to obtain a more handable description of the same system. The initial space-time coordinates are called **microscopic**, while the new ones, those well suited for the description of kinetic or hydrodynamical systems, are called **macroscopic**.

Boltzmann equation It is an integro-differential kinetic equation for the one particle distribution function in the classical phase space (see Eq. (64) below). It arises in some physical regimes, namely for a rarefied gas and for a weakly interacting quantum dense gas.

Uehling–Uhlenbeck equation It is a Boltzmann type equation taking into account corrections due to the Bose–Einstein or the Fermi–Dirac statistics (see Eq. (69) below). It holds in the weak-coupling limit.

Fokker–Planck–Landau equation It is a kinetic equation diffusive in velocity (see Eq. (28) below). It arises in the context of a weakly interacting classical dense gas.

Hydrodynamical equations They are evolution equations for macroscopic quantities like density, mean velocity, temperature, and so on.

Low density limit Sometimes called Boltzmann–Grad limit, it is a scaling limit in which the density is vanishing. Applied to classical particle systems it gives the Boltzmann equation.

Weak coupling limit It is a scaling limit in which the density is constant but the interaction vanishes suitably. Applied to a quantum particle systems it gives a Boltzmann equation. Applied to a classical particle systems it gives the Fokker–Planck–Landau equation.

Hydrodynamic limit In this scaling we simply pass from micro to macro variables. We look at the behavior of suitable mean values, which are functions of the space and the time. We expect them to behave in a hydrodynamical way, namely to satisfy a set of hydrodynamical equations.

Wigner transform It is the description of a quantum state as a function in the classical phase space.

Definition of the Subject

Many interesting systems in Physics and Applied Sciences are constituted by a large number of identical components so that they are difficult to analyze from a mathematical point of view. On the other hand, quite often, we are not interested in a detailed description of the system but rather to his global behavior. Therefore it is necessary to look for all procedures leading to simplified models, retaining all the interesting features of the original system, cutting away unnecessary informations. This is exactly the methodology of the Statistical Mechanics and the Kinetic Theory when dealing with large particle systems.

In this contribution we want to approach a problem of this type, namely we try to outline the difficulties in making rigorous the limiting procedure leading the microscopic description of a large particle system (based on the fundamental laws like Newton or Schrödinger equations) to a more handable kinetic or hydrodynamic picture. Such a transition depends on the space-time scale we choose to describe the phenomenon. To fix the ideas we start by considering a system constituted by N identical classical point particles of unitary mass. The Newton equations are:

$$\frac{d^2}{d\tau^2} q_i = \sum_{\substack{j=1\dots N:\\ j\neq i}} F(q_i - q_j) \,, \tag{1}$$

where $\{q_1 \dots q_N\}$, $q_i \in \mathbb{R}^3$ are the position of the particles and τ is the time. Here $F = -\nabla\phi$ denotes the interparticle (conservative) force and ϕ the two-body interaction potential.

We now introduce a small parameter $\varepsilon > 0$ expressing the ratio between the macroscopic and the microscopic scales. For instance ε could denote the inverse ratio of a typical distance between two molecules of a gas, measured in microscopic unities, and the same distance measured in meters. It can be as well the ratio between typical macroscopic and microscopic times. We now introduce the new variables

$$x = \varepsilon q \qquad t = \varepsilon\tau \,.$$

Since we are interested in the macroscopic properties of the system, namely in the evolution of macroscopic quantities that are those varying on the (x, t) scale and hence almost constant in the (q, τ) scale, it is natural to rescale also Eq. (1). We write

$$\frac{d^2}{dt^2} x_i = \frac{1}{\varepsilon} \sum_{\substack{j=1\dots N:\\ j\neq i}} F\left(\frac{x_i - x_j}{\varepsilon}\right) . \tag{2}$$

Up to now we did nothing else than an innocent change of variables. However to obtain a non trivial description, N has to diverge when $\varepsilon \to 0$, and we have to specify how. Also additional hypotheses on the strength of the interaction, according to the physical situation at hand, may be necessary. There are various possible scaling limits we shall discuss below.

- *The hydrodynamic limit*
  Here the system is dense, namely $N = O(\varepsilon^{-3})$. For all $x \in \Omega \subset \mathbb{R}^3$, we consider a small box Δ around x. In Δ there are a very large number of particles. We denote the initial density, mean velocity and temperature by $(\rho_0(x), u_0(x), T_0(x))$. At a positive time t the particles starts to interact. For a microscopic time τ the particles in Δ interact practically among themselves only (supposing a short range interaction) and, if an ergodic property would hold, we expect that they reach a thermal equilibrium state. We remind that $t = \varepsilon\tau$ is essentially vanishing, therefore we expect that at macroscopic time $t \approx 0$ the particle are distributed according to the Gibbs measure characterized by the parameters $(\rho_0(x), u_0(x), T_0(x))$, being mass momentum and energy locally conserved. On a larger time scale ($t > 0$) mass momentum and energy are not locally conserved anymore. If a local equilibrium structure is preserved, the parameters of the equilibria $(\rho(x, t), u(x, t), T(x, t))$ will evolve according to five equations which are the exactly the well known Euler equations for compressible gas.
  This complex mechanism is very far to be proven rigorously, even for short times and initial regular data. A heuristic derivation of the Euler equations in terms of particle dynamics was first given by Morrey [35]. We address the reader to [39] and [19] for further references and comments.
- *The low-density (Boltzmann–Grad) limit*
  Now we are dealing with a rarefied gas assuming $N = O(\varepsilon^{-2})$. This is the reason why this scaling is called low-density. We assume also, just for simplicity, that the interaction range of ϕ is one. Consider now a given test particle. If the gas is more or less homogeneous, the number of particles interacting with this particle in a given (macroscopic) time interval is $O(N\varepsilon^2) = O(1)$. Therefore we expect a finite number of collisions per unit time. Moreover each collision is almost instantaneous so that each particle undergoes a jump process in velocity.
  We also assume that initially the particles are independently distributed according to the one-particle distribution on the phase space $f_0(x, v)$ (v is the velocity). Of course the dynamics creates correlations so that, strictly speaking, we expect that the particles, for every positive time t, are not independently distributed anymore. However the probability that a given pair of particles interact is vanishing in the limit $\varepsilon \to 0$. Therefore the statistical independence (called propagation of chaos) is recovered in the same limit. In other words we expect, for the time evolved distribution $f(x, v, t)$, a nonlinear evolution equation, which is the celebrated Boltzmann equation discovered in 1872. This scaling limit, often called the Boltzmann–Grad limit [25], has been proved rigorously by Lanford [31] for a short time

interval (see also [28] for a special case in which the limit holds globally in time). We address the reader to refs. [14,39] and [19] for additional comments and references.

- *The weak-coupling limit*
  Now the gas is dense because we assume $N = O(\varepsilon^{-3})$, however the particles are weakly interacting. This condition is expressed by rescaling the interaction potential by setting

$$\phi \to \sqrt{\varepsilon}\phi \,.$$

After that the equation of motion (in macroscopic variables) become:

$$\frac{d^2}{dt^2}x_i = -\frac{1}{\sqrt{\varepsilon}} \sum_{\substack{j=1\dots N:\\ j\neq i}} \nabla\phi\left(\frac{x_i - x_j}{\varepsilon}\right) . \qquad (3)$$

To have a vague idea of the behavior of a test particle in this limit, we observe first that the velocity change due to a single collision is

$$\Delta v \approx F\Delta t \approx O(\varepsilon/\sqrt{\varepsilon}) = O(\sqrt{\varepsilon}) \,.$$

Here $\Delta t = O(\varepsilon)$ is the time in which the collision takes place, ϕ is assumed short-range in microscopic variables and then $1/\sqrt{\varepsilon}$ is the order of magnitude of the force F.
The number of collisions per unit time is $\varepsilon^2\varepsilon^{-3} = \varepsilon^{-1}$, so that

$$\sum |\Delta v|^2 = O(1) \,.$$

Therefore, provided that the propagation of chaos is ensured, $v = v(t)$ is expected to be not absolutely continuous in the limit and the one-particle distribution function to satisfy a diffusion equation in velocity, the Landau–Fokker–Planck equation.
The statistical independence at a positive time t is expected because the effect of the interaction between two given particle is going to vanish in the limit $\varepsilon \to 0$. Note that the physical meaning of the propagation of chaos here is quite different from that arising in the contest of the Boltzmann equation. Here two particles can interact but the effect of the collision is small, while in a low-density regime the effect of a collision between two given particles is large but quite unlikely.
The above scalings make sense as well for quantum systems. What we expect in this case? Notice that the transition from micro to macro variables increase the frequency of the quantum oscillations so that we are always in presence of a semiclassical limit as regards the kinetic energy part of the Hamiltonian. However the potential varies on the same scale of the oscillations. As a consequence we face the following scenario. The hydrodynamic limit of quantum systems yields the Euler equations for the equilibria parameters ρ, u, T. The only difference with the classical case is that the relationship between the pressure and the hydrodynamical fields, is that dictated by the Quantum Statistical Mechanics. Indeed the local equilibrium is achieved at a microscopic level and this is the only quantum effect we expect to survive in the limit.
The situation is conceptually similar for the low density limit: here we have the classical Boltzmann equation for the one-particle distribution function. The only quantum macroscopic effect is that the cross-section must be computed in terms of the quantum scattering process (see [6]).

Introduction

The hydrodynamical and the low density limits are relatively more popular then the weak-coupling limit, which will be, for that reason, the object of the present contribution. More precisely we analyze large classical and quantum particle systems in the weak-coupling regime. We show how we expect such systems to be well described by the Landau–Fokker–Planck and the Boltzmann equations, for the classical and the quantum case respectively. We also describe the effects of the Fermi–Dirac and Bose–Einstein statistics.

Unfortunately our arguments we present here are largely formal because the rigorous results we know up to now, are few. A rigorous proof of the validity of the weak-coupling limit, or a rigorous derivation of the corresponding kinetic equations, is a challenging and still open problem.

In Sect. “Weak-Coupling Limit for Classical Systems” we introduce and discuss the problem for classical systems, presenting a formal proof of the validity of the Landau equation.

In Sect. “Weak-Coupling Limit for Quantum Systems” we pass to analyze the case of a quantum system, neglecting the correlations due to the statistics. The case of Bosons and Fermions is discussed in Sect. “Weak-Coupling Limit in the Bose–Einstein and Fermi–Dirac Statistics”.

Finally, in Sect. “Weak-Coupling Limit for a Single Particle: the Linear Theory” we briefly introduce and discuss the corresponding linear problems, namely the behavior of a single (classical or quantum) particle in a random distribution of obstacles.

Section "Future Directions" is devoted to concluding remarks.

Weak-Coupling Limit for Classical Systems

We consider a classical system of N identical particles of unitary mass. Rescaled positions and velocities are denoted by $x_1 \dots x_N$ and $v_1 \dots v_N$. After having also rescaled the potential in the following way

$$\phi \to \sqrt{\varepsilon}\phi \, ,$$

the Newton equations reads as

$$\frac{\mathrm{d}}{\mathrm{d}t}x_i = v_i \qquad \frac{\mathrm{d}}{\mathrm{d}t}v_i = -\frac{1}{\sqrt{\varepsilon}} \sum_{\substack{j=1\dots N: \\ j\neq i}} \nabla\phi\left(\frac{x_i - x_j}{\varepsilon}\right) . \quad (4)$$

We also assume that $N = O(\varepsilon^{-3})$, namely the density is $O(1)$.

We are interested in the statistical behavior of the system so that we introduce a probability distribution on the phase space of the system, $W^N = W^N(X_N, V_N)$, that is the state at time zero. Here (X_N, V_N) denote the set of positions and velocities:

$$X_N = (x_1, \dots, x_N) \quad V_N = (v_1, \dots, v_N) \, .$$

Then from Eq. (4) we obtain the following Liouville equation

$$(\partial_t + V_N \cdot \nabla_N) W^N(X_N, V_N) = \frac{1}{\sqrt{\varepsilon}} \left(T^\varepsilon_N W^N\right)(X_N, V_N) \quad (5)$$

where $V_N \cdot \nabla_N = \sum_{i=1}^N v_i \cdot \nabla_{x_i}$ and $(\partial_t + V_N \cdot \nabla_N)$ is the usual free stream operator. Also, we have introduced the operator

$$\left(T^\varepsilon_N W^N\right)(X_N, V_N) = \sum_{0<k<\ell\leq N} \left(T^\varepsilon_{k,\ell} W^N\right)(X_N, V_N) , \quad (6)$$

with

$$T^\varepsilon_{k,\ell} W^N = \nabla\phi\left(\frac{x_k - x_\ell}{\varepsilon}\right) \cdot (\nabla_{v_k} - \nabla_{v_\ell}) W^N \, . \qquad (7)$$

To investigate the limit $\varepsilon \to 0\,,\ N = \varepsilon^{-3}$, it is convenient to introduce the BBKGY hierarchy (from Bogoliubov, Born, Green, Kirkwood and Yvon, see e. g. [2] and [19]), for the j-particle distributions defined as

$$f^N_j(X_j, V_j) = \int \mathrm{d}x_{j+1} \dots \int \mathrm{d}x_N \int \mathrm{d}v_{j+1} \dots \int \mathrm{d}v_N \\ W^N(X_j, x_{j+1} \dots x_N, V_j, v_{j+1} \dots v_N) \quad (8)$$

for $j = 1, \dots, N-1$. Obviously, we set $f^N_N = W^N$.

From now on we shall suppose that, due to the fact that the particles are identical, the function W^N and f^N_j which we have introduced are all symmetric in the exchange of particles.

A partial integration of the Liouville equation (5) and standard manipulations give us the following hierarchy of equations: (for $1 \leq j \leq N$):

$$\left(\partial_t + \sum_{k=1}^{j} v_k \cdot \nabla_k\right) f^N_j = \frac{1}{\sqrt{\varepsilon}} T^\varepsilon_j f^N_j + \frac{N-j}{\sqrt{\varepsilon}} C^\varepsilon_{j+1} f^N_{j+1} . \quad (9)$$

The operator C^ε_{j+1} is defined as:

$$C^\varepsilon_{j+1} = \sum_{k=1}^{j} C^\varepsilon_{k,j+1} \, , \qquad (10)$$

and

$$C^\varepsilon_{k,j+1} f_{j+1}(x_1 \dots x_j, v_1 \dots v_j) \\ = -\int \mathrm{d}x_{j+1} \int \mathrm{d}v_{j+1} F\left(\frac{x_k - x_\ell}{\varepsilon}\right) \qquad (11) \\ \cdot \nabla_{v_k} f_{j+1}(x_1, x_2, \dots, x_{j+1}, v_1, \dots, v_{j+1}) \, .$$

$C^\varepsilon_{k,j+1}$ describes the "collision" of particle k, belonging to the j-particle subsystem, with a particle outside the subsystem, conventionally denoted by the number $j+1$ (this numbering uses the fact that all particles are identical). The total operator C^ε_{j+1} takes into account all such collisions. The dynamics of the j-particle subsystem is governed by three effects: the free-stream operator, the "recollisions", i. e. the collisions "inside" the subsystem, given by the T term, and the "creations", i. e. the collisions with particles "outside" the subsystem, given by the C term.

We finally fix the initial value $\{f^0_j\}^N_{j=1}$ of the solution $\{f^N_j(t)\}^N_{j=1}$ assuming that $\{f^0_j\}^N_{j=1}$ is factorized, that is, for all $j = 1, \dots N$

$$f^0_j = f_0^{\otimes j} \, , \qquad (12)$$

where f_0 is a given one-particle distribution function. This means that any pair of particles are statistically uncorrelated at time zero. Of course such a statistical independence is destroyed at time $t > 0$ and Eq. (9) shows that the time evolution of f^N_1 is determined by the knowledge of f^N_2 which turns out to be dependent on f^N_3 and so on. However, since the interaction between two given particle is going to vanish in the limit $\varepsilon \to 0$, we can hope that such statistical independence, namely the factorization property (12), is recovered in the same limit.

Therefore we expect that in the limit $\varepsilon \to 0$ the one-particle distribution function f^N_1 converges to the solution of a suitable nonlinear kinetic equation f which we are going to investigate.

If we expand $f_j^N(t)$ as a perturbation of the free flow $S(t)$ defined as

$$(S(t)f_j)(X_j, V_j) = f_j(X_j - V_j t, V_j)\,, \tag{13}$$

we find

$$\begin{aligned} f_j^N(t) = S(t)f_j^0 + \frac{N-j}{\sqrt{\varepsilon}} \int_0^t S(t-t_1) C_{j+1}^{\varepsilon} f_{j+1}^N(t_1) \mathrm{d}t_1 \\ + \frac{1}{\sqrt{\varepsilon}} \int_0^t S(t-t_1) T_j^{\varepsilon} f_j^N(t_1) \mathrm{d}t_1\,. \end{aligned} \tag{14}$$

We now try to keep informations on the limit behavior of $f_j^N(t)$. Assuming for the moment that the time evolved j-particle distributions $f_j^N(t)$ are smooth (in the sense that the derivatives are uniformly bounded in ε), then

$$\begin{aligned} C_{j+1}^{\varepsilon} f_{j+1}^N(X_j, V_j, t_1) = -\varepsilon^3 \sum_{k=1}^{j} \int \mathrm{d}r \int \mathrm{d}v_{j+1} F(r) \\ \cdot \nabla_{v_k} f_{j+1}(X_j, x_k - \varepsilon r, V_j, v_{j+1}, t_1)\,. \end{aligned} \tag{15}$$

Assuming now, quite reasonably, that

$$\int \mathrm{d}r F(r) = 0\,, \tag{16}$$

we find that

$$C_{j+1}^{\varepsilon} f_{j+1}^N(X_j, V_j, t_1) = O(\varepsilon^4)$$

provided that $D_v^2 f_{j+1}^N$ is uniformly bounded. Since

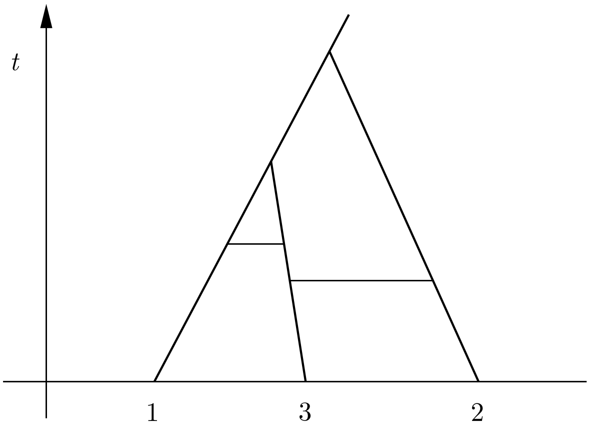
$$\frac{N-j}{\sqrt{\varepsilon}} = O(\varepsilon^{\frac{7}{2}})$$

we see that the second term in the right hand side of (14) does not give any contribution in the limit. Moreover

$$\begin{aligned} \int_0^t S(t-t_1) T_j^{\varepsilon} f_j^N(t_1) \mathrm{d}t_1 = \sum_{i \neq k} \int_0^t \mathrm{d}t_1 \\ \cdot F\left(\frac{(x_i - x_k) - (v_i - v_k)(t - t_1)}{\varepsilon} \right) g(X_j, V_j, t_1)\,, \end{aligned} \tag{17}$$

where g is a smooth function.

Obviously the above time integral is $O(\varepsilon)$ so that also the last term in the right hand side of (14) does not give any contribution in the limit. Then we are facing the alternative: either the limit is trivial or the time evolved distributions are not smooth. However we believe that the limit is not trivial (actually we expect to get a diffusion equation, according to the previous discussion) and a rigorous proof of this fact seems problematic.



Scaling Limits of Large Systems of Non-linear Partial Differential Equations, Figure 1
The collision sequence (1, 2), (1, 3), (1, 3), (2, 3)

The difficulty in obtaining a-priori estimates induce us to exploit the full series expansion of the solution, namely

$$\begin{aligned} & f_1^N(t) \\ & = \sum_{n \geq 0} \sum_{G_n} K(G_n) \int_0^t \mathrm{d}t_1 \int_0^{t_1} \mathrm{d}t_2 \ldots \int_0^{t_{n-1}} \mathrm{d}t_n \\ & \quad \cdot \left[S(t-t_1) O_1 S(t_1 - t_2) \ldots O_n S(t_n) \right] f_m^0\,. \end{aligned} \tag{18}$$

Here O_j is either an operator C or T expressing a creation of a new particle or a recollision between two particles respectively. G_n is a graph namely a sequence of indices

$$(r_1, l_1), (r_2, l_2), \ldots (r_n, l_n)$$

where (r_j, l_j), $r_j < l_j$ are the pair of indices of the particles involved in the interaction at time t_j. The number of particles created in the process is $m - 1$. It is convenient to represent the generic graph in the following way, as in Fig. 1.

The legs of the graph denotes the particles and the nodes the creation of new particles (operators C). Recollisions (operators T) are represented by horizontal link. For instance the graph in figure is $(1, 2)$, $(1, 3)$, $(1, 3)$, $(2, 3)$, $m = 3$ and the integrand in Eq. (18) is in this case

$$\begin{aligned} & \left[S(t-t_1) C_{1,2} S(t_1 - t_2) C_{1,3} S(t_2 - t_3) T_{1,3} \right. \\ & \quad \left. \cdot S(t_3 - t_4) T_{2,3} S(t_4) \right] f_3^0\,. \end{aligned} \tag{19}$$

Note that the knowledge of the graph determines completely the sequence of operators in the right hand side of (20).

Finally the factor $K(G_n)$ takes into account the divergences:

$$K(G_n) = O\left(\varepsilon^{-\frac{n}{2}}\varepsilon^{-3(m-1)}\right) . \tag{20}$$

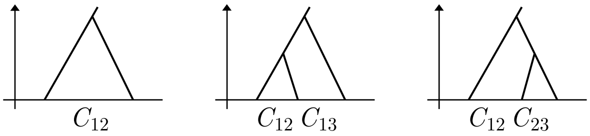
We are not able to analyze the asymptotic behavior of each term of the expansion (18) however we can compute the limit for $\varepsilon \to 0$ of the few terms up to the second order (in time). We have:

$$\begin{aligned} g^N(x_1, v_1, t) &= f^0(x_1 - v_1 t, v_1) + \frac{N-1}{\sqrt{\varepsilon}} \int_0^t S(t-t_1) \\ &\cdot C^{\varepsilon}_{1,2} S(t_1) f^0_2 \mathrm{d}t_1 + \frac{(N-1)}{\varepsilon}\frac{(N-2)}{\varepsilon} \sum_{j=1,2} \int_0^{t_1} \mathrm{d}t_2 \\ &\cdot S(t-t_1) C^{\varepsilon}_{1,2} S(t_1-t_2) C^{\varepsilon}_{j,3} S(t_2) f^0_3 + \frac{N-1}{\varepsilon} \int_0^t \mathrm{d}t_1 \\ &\cdot \int_0^{t_1} \mathrm{d}t_2\, S(t-t_1) C^{\varepsilon}_{1,2} S(t_1-t_2) T^{\varepsilon}_{1,2} S(t_2) f^0_2 . \end{aligned} \tag{21}$$

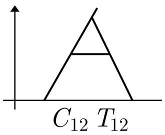
Here the right hand side of (21) defines g^N.

The second and third term in (21) corresponding to the graphs shown in Fig. 2 are indeed vanishing as follows by the use of the previous arguments.

The most interesting term is the last one, shown in Fig. 3.



Scaling Limits of Large Systems of Non-linear Partial Differential Equations, Figure 2
The vanishing terms in the expansion of g^N



Scaling Limits of Large Systems of Non-linear Partial Differential Equations, Figure 3
The first non vanishing terms in the expansion of g^N

To handle this term we denote by $w = v_1 - v_2$ the relative velocity and note that, for a given function u:

$$\begin{aligned} & S(t_1-t_2) T^{\varepsilon}_{1,2} u(x_1, x_2, v_1, v_2) \\ &= -F\left(\frac{(x_1-x_2) - w(t_1-t_2)}{\varepsilon}\right) \cdot [(\nabla_{v_1} - \nabla_{v_2})u] \\ &\quad \cdot (x_1 - v_1(t_1-t_2), x_2 - v_2(t_1-t_2), v_1, v_2) \\ &= -F\left(\frac{(x_1-x_2) - w(t-t_1)}{\varepsilon}\right) \cdot (\nabla_{v_1} - \nabla_{v_2} + (t_1-t_2) \\ &\quad \cdot (\nabla_{x_1} - \nabla_{x_2})) S(t_1-t_2) u(x_1, x_2, v_1, v_2) . \end{aligned} \tag{22}$$

Therefore the last term in the r.h.s. of (22) is

$$\begin{aligned} & \frac{N-1}{\varepsilon} \int_0^t \mathrm{d}t_1\, S(t-t_1) \int_0^{t_1} \mathrm{d}t_2 \int \mathrm{d}x_2 \int \mathrm{d}v_2 \\ & F\left(\frac{x_1-x_2}{\varepsilon}\right) \cdot \nabla_{v_1} F\left(\frac{(x_1-x_2) - w(t_1-t_2)}{\varepsilon}\right) \\ &\quad \cdot (\nabla_{v_1} - \nabla_{v_2} + (t_1-t_2)(\nabla_{x_1} - \nabla_{x_2})) \\ &\quad \cdot S(t_1) f^0_2(x_1, x_2, v_1, v_2) . \end{aligned} \tag{23}$$

Setting now $r = \frac{x_1-x_2}{\varepsilon}$ and $s = \frac{t_1-t_2}{\varepsilon}$ then

$$\begin{aligned} & g^N(x_1, v_1, t) = (N-1)\varepsilon^3 \int_0^t \mathrm{d}t_1 \int_0^{\frac{t_1}{\varepsilon}} \mathrm{d}s \int \mathrm{d}r \\ &\quad \cdot \int \mathrm{d}v_2\, F(r) \cdot \nabla_{v_1} \\ & F(r - ws) \cdot (\nabla_{v_1} - \nabla_{v_2} + \varepsilon s(\nabla_{x_1} - \nabla_{x_2})) \\ &\quad \cdot S(t_1 - \varepsilon s) f^0_2(x_1, x_2, v_1, v_2) + O(\sqrt{\varepsilon}) . \end{aligned} \tag{24}$$

The formal limit is of (21) is

$$\begin{aligned} g(x_1, v_1, t) = \int_0^t \mathrm{d}t_1 \int \mathrm{d}v_2\, S(t-t_1) \nabla_{v_1} a(v_1 - v_2) \\ \cdot (\nabla_{v_1} - \nabla_{v_2})\ S(t_1) f^0_2 , \end{aligned} \tag{25}$$

where (using $F(r) = -F(-r)$) the matrix a is given by:

$$\begin{aligned} a(w) &= \int \mathrm{d}r \int_0^{+\infty} \mathrm{d}s\, F(r) \otimes F(r - ws) = \frac{1}{2} \int \mathrm{d}r \\ &\cdot \int_{-\infty}^{+\infty} \mathrm{d}s\, F(r) \otimes F(r - ws) \frac{1}{2} \frac{1}{(2\pi)^3} \int_{-\infty}^{+\infty} \mathrm{d}s \\ &\cdot \int \mathrm{d}k\, k \otimes k\, \hat{\phi}(k)^2 \mathrm{e}^{i(w \cdot k)s} = \frac{1}{(8\pi)^2} \int \mathrm{d}k\, k \otimes k\, \hat{\phi}(k)^2 \\ &\cdot \delta(w \cdot k) = \frac{A}{|w|^3} (|w|^2 Id - w \otimes w), \end{aligned} \tag{26}$$

where $\hat{\phi}(k) = \int e^{-il\cdot x}\phi(x)\mathrm{d}x$. Here the interaction potential ϕ has been assumed spherically symmetric and:

$$A = \frac{1}{8\pi}\int_0^{+\infty} \mathrm{d}r\, r^3 \hat{\phi}(r)^2\,. \tag{27}$$

Looking at Eq. (25) we are led to introduce the following nonlinear equation:

$$(\partial_t + v\cdot\nabla_x)f = Q_L(f, f) \tag{28}$$

with the collision operator Q_L given by:

$$Q_L(f,f)(v) = \int \mathrm{d}v_1 \cdot \nabla_v\Big[a(v-v_1)\,(\nabla_v - \nabla_{v_1})\; f(v)f(v_1)\Big]. \tag{29}$$

Here x plays the role of a parameter and hence his dependence is omitted.

Equation (28) is called the Fokker–Planck–Landau equation (Landau equation in the sequel) and has been introduced by Landau in the study of a dense, weakly interacting gas (see [32]).

From (28) we obtain the following (infinite) hierarchy of equations

$$(\partial_t + V_j\cdot\nabla_{X_j})f_j = C_{j+1}f_{j+1} \tag{30}$$

for the quantities:

$$f_j(t) = f(t)^{\otimes j} \tag{31}$$

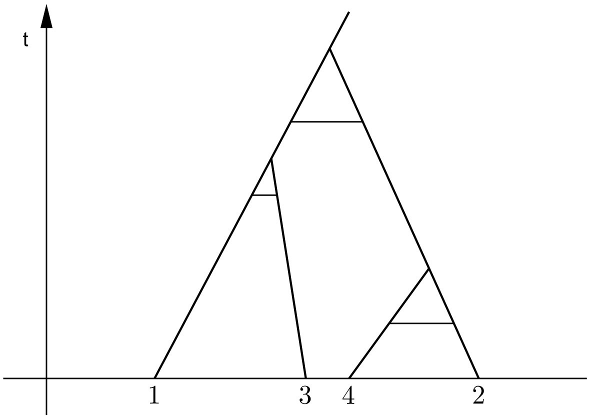
where $f(t)$ solves Eq. (28). Accordingly $C_{j+1} = \sum_k C_{k,j+1}$ where

$$\begin{aligned} &C_{k,j+1}f_{j+1}(x_1\dots x_j, v_1\dots v_j)\\ &= \prod_{r\neq k} f(x_r, v_r)Q_L(f,f)(x_k, v_k)\,. \end{aligned} \tag{32}$$

Therefore f has the following series expansion representation

$$\begin{aligned} f(t) = \sum_{n\geq 0}\int_0^t \mathrm{d}t_1 \int_0^{t_1}\mathrm{d}t_2 \dots \int_0^{t_{n-1}} \mathrm{d}t_n\\ \cdot\big[S(t-t_1)C_2S(t_1-t_2)C_3\dots C_{n-1}S(t_n)\big]f^0_{n+1}\,. \end{aligned} \tag{33}$$

The previous calculation shows the formal convergence of g^N to the term with $n = 1$ of the expansion (33), namely we have an agreement between the particle system (18) and the solution to the Landau equation (33) at least up to the first order in time (second order in the potential). Although the above arguments can be made rigorous, under suitable assumption on the initial condition f_0 and the potential ϕ, it seems difficult to show the convergence of the whole series. On the other hand it is clear that the graphs which should contribute in the limit are those formed by a collision-recollision sequence, like that shown in Fig. 4. For those terms it is probably possible to show the convergence. For instance the case in figure has the asymptotics



Scaling Limits of Large Systems of Non-linear Partial Differential Equations, Figure 4
The collision-recollision sequence $(1,2)$, $(1,2)$, $(1,3)$, $(1,3)$, $(2,4)$, $(2,4)$

$$\begin{aligned} \int_0^t \mathrm{d}t_1 \int_0^{t_1}\mathrm{d}t_2\int_0^{t_2}\mathrm{d}t_3\big[S(t-t_1)C_{1,2}S(t_1-t_2)\\ \cdot C_{1,3}S(t_2-t_3)C_{2,4}S(t_3)\big]f^0_4\,. \end{aligned} \tag{34}$$

However the proof that all other graphs are vanishing in the limit is not easy. Even more difficult is a uniform control of the series expansion (18), even for short times. As we shall see in the next section, something more can be obtained for quantum systems under the same scaling limit.

Some comments are in order.

In the present section we showed how the Landau equation is expected to be derived in the weak-coupling limit from a particle system. In doing this we essentially followed the monograph [2]. This is not however the usual way in which the Landau equation is introduced in the literature. Indeed it is usually recovered from the Boltzmann equation. when the density increases and the grazing collisions become dominant. In particular, the case of the homogeneous Boltzmann equation has been investigated in [1,24,44] (see also [15] for the Coulomb potential case, and see the general survey [43]). In [1] the authors show that, under suitable assumptions on the cross-section, the diffusion Fokker–Planck–Landau equation (28) can indeed be derived. The diffusion operator is the form (29) but with a matrix a given by

$$a(w) = \alpha(|w|)(|w|^2 Id - w \otimes w) ,$$

with α smooth function. Next in [24] and [44] steps forward were performed to arrive to cover the case $\alpha(|w|) \approx 1/|w|^\nu$ for small $|w|$, with $\nu < 1$. We remark that the diffusion coefficient found in this way is different from that derived here in (26).

We conclude by observing that the Landau equation describes a genuine kinetic evolution. Mass, momentum and energy are conserved, while the kinetic entropy is decreasing. The equilibrium is Maxwellian. The difficulties in the validation problem are certainly related to the transition from a reversible (hamiltonian) dynamics to an irreversible one, as for the Boltzmann equation. Here we tried to compare the two series expansions as in the Lanford's validation proof for the Boltzmann equation. This is a sort of Cauchy–Kowalevski brute force argument which works, as well, for negative times. It is clear that, in this way, we cannot go beyond a short time result and even this seems not trivial at all. Other approaches making use of the diffusive nature of the motion of a test particle are, at moment, absent. However the weak-coupling limit of a single particle in a random distribution of scatterers is well understood as we shall discuss later in Sect. "The Weak Coupling Limit for a Single Particle: the Linear Theory".

Weak-Coupling Limit for Quantum Systems

We consider the quantum analog of the system considered in Sect. "Introduction", namely N identical quantum particles with unitary mass in $\mathbb{R}^3$. In the present section the statistical nature of the particles will be ignored.

The interaction between particles is still a two-body potential ϕ so that the total potential energy is taken as

$$U(x_1 \dots x_N) = \sum_{i<j} \phi(x_i - x_j) . \tag{35}$$

The associated Schrödinger equation reads

$$i\partial_t \Psi(X_N, t) = -\tfrac{1}{2}\Delta_N \Psi(X_N, t) + U(X_N)\Psi(X_N, t), \tag{36}$$

where $\Delta_N = \sum_{i=1}^N \Delta_i$, Δ_i is the Laplacian with respect to the x_i variables, $X_N = (x_1, \dots, x_N)$ and $\hbar$ is normalized to unity.

As for the classical system considered in Sect." Introduction" we rescale the equation and the potential by

$$x \to \varepsilon x , \qquad t \to \varepsilon t , \qquad \phi \to \sqrt{\varepsilon}\phi . \tag{37}$$

The resulting equation is,

$$i\varepsilon\partial_t \Psi^\varepsilon(X_N, t) = -\frac{\varepsilon^2}{2}\Delta_N \Psi^\varepsilon(X_N, t) + U_\varepsilon(X_N)\,\Psi^\varepsilon(X_N, t), \tag{38}$$

where:

$$U_\varepsilon(x_1 \dots x_N) = \sqrt{\varepsilon}\sum_{i<j} \phi\left(\frac{x_i - x_j}{\varepsilon}\right) . \tag{39}$$

We want to analyze the limit $\varepsilon \to 0$ in the above equations, when $N = \varepsilon^{-3}$.

Note that this limit looks, at a first sight, similar to a semiclassical (or high frequency) limit. It is not so: indeed the potential varies on the same scale of the typical oscillations of the wave functions so that the scattering process is a genuine quantum process. Obviously, due to the oscillations, we do not expect that the wave function does converge to something in the limit. The right quantity to look at was introduced by Wigner in 1922 [45] to deal with kinetic problems. It is called the Wigner transform (of Ψ^ε) and is defined as

$$\begin{aligned} W^N(X_N, V_N) = \left(\frac{1}{2\pi}\right)^{3N} \int \mathrm{d}Y_N\, \mathrm{e}^{iY_N \cdot V_N}\overline{\Psi}^\varepsilon \\ \cdot (X_N + \frac{\varepsilon}{2}Y_N)\,\Psi^\varepsilon(X_N - \frac{\varepsilon}{2}Y_N) . \end{aligned} \tag{40}$$

The Wigner transform W^N satisfies a transport-like equation, completely equivalent to the Schrödinger equation:

$$(\partial_t + V_N \cdot \nabla_N)W^N(X_N, V_N) = \frac{1}{\sqrt{\varepsilon}}(T_N^\varepsilon W^N)(X_N, V_N). \tag{41}$$

The operator T_N^ε on the right-hand-side of (38) plays the same role of the classical operator denoted with the same symbol in Sect. "Introduction". It is

$$(T_N^\varepsilon W^N)(X_N, V_N) = \sum_{0<k<\ell\leq N} (T_{k,\ell}^\varepsilon W^N)(X_N, V_N), \tag{42}$$

where each $T_{k,\ell}^\varepsilon$ describes the interaction of particle k with particle ℓ

$$\begin{aligned} (T_{k,\ell}^\varepsilon W^N)(X_N, V_N) = \frac{1}{i}\Bigg(\frac{1}{(2\pi)^{3N}} \int \mathrm{d}Y_N \mathrm{d}V_N' \\ \cdot \mathrm{e}^{iY_N\cdot(V_N - V_N')}\left[\phi\left(\frac{x_k - x_\ell}{\varepsilon} - \frac{y_k - y_\ell}{2}\right)\right) \\ -\phi\left(\frac{x_k - x_\ell}{\varepsilon} + \frac{y_k - y_\ell}{2}\right)\right] W^N(X_N, V_N') . \end{aligned} \tag{43}$$

Equivalently, we may write

$$(T^{\varepsilon}_{k,\ell}W^N)(X_N,V_N) = -i\sum_{\sigma=\pm1}\sigma\int\frac{\mathrm{d}h}{(2\pi)^3}\,\hat{\phi}(h)\cdot \mathrm{e}^{i\frac{h}{\varepsilon}(x_k-x_\ell)}W^N(x_1,\dots,x_N;v_1,\dots,v_k-\sigma\frac{h}{2},\dots,v_\ell+\sigma\frac{h}{2},\dots,v_N)\,. \quad (44)$$

Note that $T^{\varepsilon}_{k,\ell}$ is a pseudodifferential operator which formally converge, at fixed ε, for $\hbar\to0$ (here $\hbar=1$) to its classical analog given by Eq. (7). Note also that in (44), "collisions" may take place between *distant* particles ($x_k\neq x_\ell$). However, such distant collisions are penalized by the highly oscillatory factor $\exp(ih(x_k-x_\ell)/\varepsilon)$. These oscillations turn out to play a crucial role throughout the analysis, and they explain why collisions tend to happen when $x_k=x_\ell$ in the limit $\varepsilon\to0$.

The formalism we have introduced is formally similar to the classical case so that we proceed as before by transforming Eq. (38) into a hierarchy of equations. We introduce the partial traces of the Wigner transform W^N, denoted by f^N_j. They are defined through the following formula, valid for $j=1,\dots,N-1$:

$$f^N_j(X_j,V_j)=\int\mathrm{d}x_{j+1}\dots\int\mathrm{d}x_N\int\mathrm{d}v_{j+1}\dots\int\mathrm{d}v_N\, W^N(X_j,x_{j+1}\dots x_N,V_j,v_{j+1}\dots v_N)\,. \quad (45)$$

Obviously, we set $f^N_N=W^N$. The function f^N_j is the kinetic object that describes the state of the j particles subsystem at time t.

As before, the wave function Ψ, as well W^N and f^N_j, are assumed to be symmetric in the exchange of particle, a property that is preserved in time.

Proceeding then as in the derivation of the BBKGY hierarchy for classical systems, we readily transform Eq. (38) into the following hierarchy:

$$(\partial_t+\sum_{k=1}^{j}v_k\cdot\nabla_k)f^N_j(X_j,V_j) = \frac{1}{\sqrt{\varepsilon}}T^{\varepsilon}_j f^N_j+\frac{N-j}{\sqrt{\varepsilon}}C^{\varepsilon}_{j+1}f^N_{j+1}\,, \quad (46)$$

where

$$C^{\varepsilon}_{j+1}=\sum_{k=1}^{j}C^{\varepsilon}_{k,j+1}\,, \quad (47)$$

and $C^{\varepsilon}_{k,j+1}$ is defined by

$$C^{\varepsilon}_{k,j+1}f^N_{j+1}(X_j,V_j) = -i\sum_{\sigma=\pm1}\sigma\int\frac{\mathrm{d}h}{(2\pi)^3}\int\mathrm{d}x_{j+1}\cdot\int\mathrm{d}v_{j+1}\,\hat{\phi}(h)\,\mathrm{e}^{i\frac{h}{\varepsilon}(x_k-x_{j+1})} f^N_{j+1}\Big(x_1,x_2,\dots,x_{j+1},v_1,\dots,v_k-\sigma\frac{h}{2},\dots,v_{j+1}+\sigma\frac{h}{2}\Big)\,. \quad (48)$$

As before the initial value $\{f^0_j\}^N_{j=1}$ is assumed completely factorized: for all $j=1,\dots,N$, we suppose

$$f^0_j=f_0^{\otimes j}\,, \quad (49)$$

where f_0 is a one-particle Wigner function, and f^0 is assumed to be a probability distribution.

In the limit $\varepsilon\to0$, we expect that the j-particle distribution function $f^N_j(t)$, that solves the hierarchy (46) with initial data (49), tends to be factorized for all times: $f^N_j(t)\sim f(t)^{\otimes j}$ (propagation of chaos).

As for the classical case, if f_{j+1} is smooth:

$$C^{\varepsilon}_{k,j+1}f^N_{j+1}(X_j,V_j) = -i\varepsilon^3\sum_{\sigma=\pm1}\sigma\int\frac{\mathrm{d}h}{(2\pi)^3}\hat{\phi}(h)\cdot\int\mathrm{d}r\int\mathrm{d}v_{j+1}\;\mathrm{e}^{ih\cdot r} f^N_{j+1}\Big(X_j,x_k-\varepsilon r,\ v_1,\dots,v_k-\sigma\frac{h}{2},\dots,v_{j+1}+\sigma\frac{h}{2}\Big) = O(\varepsilon^4)\,. \quad (50)$$

Indeed, setting $\varepsilon=0$ in the integrand, the integration over r produces $\delta(h)$. As a consequence the integrand is independent on σ and the sum vanishes. Therefore the integral is $O(\varepsilon)$. Also

$$\frac{1}{\sqrt{\varepsilon}}\int_0^t\mathrm{d}t_1\,S(t-t_1)T_{r,k}f^N_j(t_1) = -i\sum_{\sigma=\pm1}\sigma\int_0^t\mathrm{d}t_1\frac{\mathrm{d}h}{(2\pi)^3}\hat{\phi}(h)\mathrm{e}^{i\frac{h}{\varepsilon}\cdot(x_r-x_k)-(v_r-v_k)(t-t_1)}\cdot f^N_j(X_j-V_j(t-t_1),V_j,t_1) \quad (51)$$

is weakly vanishing, by a stationary phase argument (see [4]). Therefore, as for the classical case, we analyze the asymptotics of the collision-recollision term (shown in Fig. 1):

$$-\frac{N-1}{\varepsilon}\int_0^t \mathrm{d}t_1 \int_0^{t_1} \mathrm{d}\tau_1 \cdot S(t-t_1)C_{1,2}S(t_1-\tau_1)T_{1,2}S(\tau_1)f_2^0 . \tag{52}$$

Explicitly it looks as:

$$\begin{aligned}&-\frac{N-1}{\varepsilon}\sum_{\sigma,\sigma'=\pm 1}\sigma\sigma'\int_0^t \mathrm{d}t_1\int_0^{t_1}\mathrm{d}\tau_1\int \mathrm{d}x_2\int \mathrm{d}v_2\\ &\cdot\int\frac{\mathrm{d}h}{(2\pi)^3}\int\frac{\mathrm{d}k}{(2\pi)^3}\hat{\phi}(h)\,\hat{\phi}(k)\,\mathrm{e}^{i\frac{h}{\varepsilon}\cdot\left(x_1-x_2-v_1(t-t_1)\right)}\\ &\cdot\mathrm{e}^{i\frac{k}{\varepsilon}\cdot\left(x_1-x_2-v_1(t-t_1)-(v_1-v_2-\sigma h)(t_1-\tau_1)\right)}\,f_2^0\Bigg(x_1-v_1 t\\ &\quad+\sigma\frac{h}{2}t_1+\sigma'\frac{k}{2}\tau_1, x_2-v_2t_1-\sigma\frac{h}{2}t_1-\sigma'\frac{k}{2}\tau_1;\\ &\quad v_1-\sigma\frac{h}{2}-\sigma'\frac{k}{2}, v_2+\sigma\frac{h}{2}+\sigma'\frac{k}{2}\Bigg) .\end{aligned} \tag{53}$$

By the change of variables:

$$t_1-\tau_1=\varepsilon s_1 , \;(\text{i.e. } \tau_1=t_1-\varepsilon s_1) , \;\xi=(h+k)/\varepsilon , \tag{54}$$

we have

$$\begin{aligned}(54) &= -(N-1)\,\varepsilon^3\sum_{\sigma,\sigma'=\pm 1}\sigma\sigma'\int_0^t \mathrm{d}t_1\int_0^{t_1/\varepsilon}\mathrm{d}s_1\\ &\cdot\int \mathrm{d}x_2\,\mathrm{d}v_2\,\frac{\mathrm{d}\xi}{(2\pi)^3}\,\frac{\mathrm{d}k}{(2\pi)^3}\,\hat{\phi}(-k+\varepsilon\xi_1)\,\hat{\phi}(k)\\ &\cdot\mathrm{e}^{i\xi\cdot\left(x_1-x_2-v_1(t-t_1)\right)}\mathrm{e}^{-is_1k\cdot(v_1-v_2-\sigma(-k+\varepsilon\xi))}f_{j+1}^0(\ldots) ,\end{aligned} \tag{55}$$

In the limit $\varepsilon \to 0$, the above formula gives the asymptotics

$$\begin{aligned}(52) \underset{\varepsilon\to 0}{\sim} &-\sum_{\sigma,\sigma'=\pm 1}\sigma\sigma'\int_0^t \mathrm{d}t_1\int \mathrm{d}v_2\,\frac{\mathrm{d}k}{(2\pi)^3}\,|\hat{\phi}(k)|^2\\ &\cdot\left(\int_0^{+\infty}\mathrm{e}^{-is_1k\cdot(v_1-v_2+\sigma k)}\mathrm{d}s_1\right)\\ &\cdot f_2^0\Bigg(x_1-v_1t-(\sigma-\sigma')\frac{k}{2}t_1,\\ &x_1-v_1(t-t_1)-v_2t_1+(\sigma-\sigma')\frac{k}{2}t_1, v_1+(\sigma-\sigma')\frac{k}{2},\\ &v_2-(\sigma-\sigma')\frac{k}{2}\Bigg) .\end{aligned} \tag{56}$$

In [4], we completely justify formula (56) and its forthcoming consequences.

Now, we turn to identifying the limiting value obtained in (56). To do so, we observe that symmetry arguments allow us to replace the integral in s by its real part:

$$\operatorname{Re}\int_0^{\infty}\mathrm{e}^{-is_1k\cdot(v_1-v_2+\sigma k)}\mathrm{d}s_1=\pi\delta(k\cdot(v_1-v_2+\sigma k)) . \tag{57}$$

Using this we realize that the contribution $\sigma = -\sigma'$ in (56) gives rise to the gain term:

$$\begin{aligned}&\int_0^t \mathrm{d}t_1\int \mathrm{d}v_2\int \mathrm{d}\omega\,B(\omega, v_1-v_2)\,f_2^0\big(x_1-v_1(t-t_1)\\ &\quad-v_1't_1, x_2-v_2(t-t_1)-v_2't_1, v_1', v_2'\big) ,\end{aligned} \tag{58}$$

where the integral in ω is on the surface of the unitary sphere in $\mathbb{R}^3$, and

$$B(\omega, v)=\frac{1}{8\pi^2}|\omega\cdot v|\,|\hat{\phi}(\omega\,(\omega\cdot v))|^2, \tag{59}$$

and the velocities v_1', v_2' are

$$v_1'=v_1-\omega(\omega\cdot(v_2-v_1)) , \quad v_2'=v_2+\omega(\omega\cdot(v_2-v_1)) .$$

Similarly, the term $\sigma = \sigma'$ in (5) yields the loss term:

$$\begin{aligned}&\int_0^t \mathrm{d}t_1\int \mathrm{d}v_2\int \mathrm{d}\omega\,B(\omega, v_1-v_2)\\ &\cdot f_2^0(x_1-v_1t, x_2-v_2(t-t), v_1, v_2) .\end{aligned} \tag{60}$$

By the same arguments used in the previous section we can conclude that the full series expansion (20) (of course for the present quantum case) agrees, up to the second order in the potential, with

$$S(t)f_0+\int_0^t \mathrm{d}t_1\,S(t-t_1)Q(S(t_1)f_0, S(t_1)f_0) \tag{61}$$

where

$$\begin{aligned}Q(f,f)&=\frac{1}{4\pi^2}\int \mathrm{d}v_1\int \mathrm{d}h\,|\hat{\phi}(h)|^2\delta((h\cdot(v-v_1+h))\\ &\cdot[f(v+h)f(v_1-h)-f(v)f(v_1)]=\int \mathrm{d}v_1\int \mathrm{d}\omega\\ &\cdot B(\omega, v-v_1)[f'f_1'-ff_1] ,\end{aligned} \tag{62}$$

with

$$\begin{aligned}&f_1=f(v_1) ,\; f'=f(v') ,\; f_1'=f(v_1') ,\\ &v'=v+h=v-\omega(\omega\cdot(v-v_1)) ,\\ &v_1'=v_1-h=v_1+\omega(\omega\cdot(v-v_1)) .\end{aligned} \tag{63}$$

In other words the kinetic equation which comes out is the Boltzmann equation with cross-section B:

$$(\partial_t + v \cdot \nabla_x) f = Q(f, f) \,, \tag{64}$$

where Q is given by Eq. (62). We note once more that the δ function in Eq. (62) expresses the energy conservation, while the momentum conservation is automatically satisfied. Note also that the cross-section B is the only quantum factor in the purely classical expression (62). It retains the quantum features of the elementary "collisions".

An important comment is in order. Why the kinetic equation for quantum systems is of Boltzmann type in contrast with the classical case where we got a diffusion? The answer is related to the asymptotics of a single scattering (see [36,37] and [8]). For quantum systems we have a finite probability of having any angle scattering, while for a classical particle, we surely have a small deviation from the free motion. Therefore a quantum particle, in this asymptotic regime, is going to perform a jump process (in velocity) rather than a diffusion.

From a mathematical view point we observe that [4] proves more than agreement up to second order. We indeed consider the subseries (of the full series expansion expressing $f_j^N(t)$) formed by *all* the collision–recollision terms (as that shown in Fig. 3). In other words, we consider the subseries of $f_j^N(t)$ given by

$$\begin{aligned}\sum_{n\geq 1} \sum_{r_1,\ldots,r_n,l_1,\ldots,l_n} \varepsilon^{-4n} \int_0^t \mathrm{d}t_1 \int_0^{t_1} \mathrm{d}\tau_1 \, S(t-t_1) C^{\varepsilon}_{r_1,l_1} \\ \cdot S(t_1-\tau_1) T^{\varepsilon}_{r_1,l_1} \int_0^{\tau_{n-1}} \mathrm{d}t_n \int_0^{t_n} \mathrm{d}\tau_n \, S(\tau_{n-1}-t_n) C^{\varepsilon}_{r_n,l_n} \\ \cdot S(t_n-\tau_n) T^{\varepsilon}_{r_n,l_n} S(\tau_n) f^0_{j+n+1} \,.\end{aligned} \tag{65}$$

Here the sum runs over all possible choices of the particles number r's and l's, namely we sum over the subset of graph of the form in Fig. 4. We establish in [4] that the subseries (65) is indeed absolutely convergent, for short times, uniformly in ε. Moreover, we prove that it approaches the corresponding complete series expansion obtained by solving iteratively the Boltzmann equation with collision operator given by Eq. (62) extending and making rigorous the above argument.

Under reasonable smoothness hypotheses on the potential and on the initial distribution, assuming in addition that $\hat{\phi}(0) = 0$, we also proved in [7], that all other terms than those considered in the subseries (65) are indeed vanishing in the limit. The condition $\hat{\phi}(0) = 0$ is probably only technical: it takes care automatically of some divergences which are difficult to deal with differently.

Unfortunately this result, even under these severe assumptions on the potential, is not yet conclusive because we are not able to show a uniform bound on the full series. Thus a mathematical justification of the quantum Boltzmann equation is a still an open and difficult problem.

It is not surprising that we know more for the quantum case in comparison with the classical problem introduced in Sect. "Introduction". Now the operators involved are differences while, for the classical case, we have to deal with derivatives. Introducing explicitly the $\hbar$ dependence, we easily realize that the results discussed in the present section are not uniform in $\hbar$.

Weak-Coupling Limit in the Bose–Einstein and Fermi–Dirac Statistics

In this section we approach the same problem as in Sect. "Weak-Coupling Limit for Classical Systems", for Bosons or Fermions, namely for particles obeying the Fermi–Dirac or Bose–Einstein statistics. As we shall see, the kinetic equation we expect to hold in the limit, is the so called Uehling and Uhlembeck equation, which is a Boltzmann type equation, with cubic collision operator. We note that the effects of the quantum correction here, enter also in the structure of the operator.

In this case, the starting point is still the rescaled Schrödinger equation (38), or the equivalent hierarchy (46). The only new point is that we cannot take a totally decorrelated initial datum as in (49). Indeed, the Fermi–Dirac or Bose–Einstein statistics yield correlations even at time zero. In this perspective, the most uncorrelated states one can introduce, and that do not violate the Fermi–Dirac or Bose–Einstein statistics, are the so called quasi-free states. They have, in terms of the Wigner formalism, the following form

$$f_j(x_1, v_1, \ldots, x_j, v_j) = \sum_{\pi \in \mathcal{P}_j} \theta^{s(\pi)} f_j^{\pi}(x_1, v_1, \ldots, x_j, v_j) \,, \tag{66}$$

where each f_j^{π} has the value

$$\begin{aligned} f_j^{\pi}(x_1, v_1, \ldots, x_j, v_j) = \int \mathrm{d}y_1 \ldots \mathrm{d}y_j \int \mathrm{d}w_1 \ldots \mathrm{d}w_j \\ \cdot \, \mathrm{e}^{i(y_1 \cdot v_1 + \cdots + y_j \cdot v_j)} \prod_{k=1}^{j} \mathrm{e}^{-\frac{i}{\varepsilon} w_k \cdot (x_k - x_{\pi(k)})} \\ \cdot \, \mathrm{e}^{-\frac{i}{2} w_k \cdot (y_k + y_{\pi(k)})} f\left(\frac{x_k + x_{\pi(k)}}{2} + \varepsilon \frac{y_k - y_{\pi(k)}}{4}, w_k \right) , \end{aligned} \tag{67}$$

and f is a given one-particle Wigner function. Here $\mathcal{P}_j$ denotes the group of all the permutations of j objects and π its generic element.

Note that the uncorrelated case treated in Sect. "Weak-Coupling Limit for Classical Systems", is recovered by the contribution due the permutation $\pi = \text{identity}$.

To verify that (66) and (67) really describe admissible states for the statistics, consider the inverse Wigner transform of (66) and (67). The density matrix obtained in this way does verify the permutation invariance required by the statistics.

Moreover the quasi-free states converge weakly to the completely factorized states as $\varepsilon \to 0$. This is physically obvious because the quantum statistics become irrelevant in the semiclassical limit. However the dynamics take place on the scale ε so that the effects of the statistics are present in the limit. Indeed it is expected that the one-particle distribution function $f_1^N(t)$ converges to the solution of the following cubic Boltzmann equation:

$$(\partial_t + v \cdot \nabla_x)\, f(x,v,t) = Q_\theta(f)(x,v,t)\,, \tag{68}$$

$$\begin{aligned} Q_\theta(f)(x,v,t) &= \int \mathrm{d}v_1\, \mathrm{d}\omega\, B_\theta(\omega, v - v_1) \\ &\cdot \left[f(x,v')f(x,v_1')(1 + 8\pi^3\theta f(x,v))(1 + 8\pi^3\theta f(x,v_1)) \right. \\ &\left. - f(x,v)f(x,v_1)(1 + 8\pi^3\theta f(x,v')(1 + 8\pi^3\theta f(x,v_1'))\right]\,, \end{aligned} \tag{69}$$

(for the notations see Eq. (63)). Here $\theta = +1$ or $\theta = -1$, for the Bose–Einstein or the Fermi–Dirac statistics respectively. Finally, B_θ is the symmetrized or antisymmetrized cross-section derived from B (see (59)) in a natural way:

$$\begin{aligned} B_\theta(\omega, v) = \frac{1}{16\pi^2}|\omega \cdot v| \\ \cdot \left[\hat{\phi}(\omega\,(\omega \cdot v)) + \theta\hat{\phi}(v - \omega\,(\omega \cdot v))\right]^2 . \end{aligned}$$

As we see, the modification of the statistics transforms the quadratic Boltzmann equation of the Maxwell–Boltzmann case, into a cubic one (fourth order terms cancel). Also, the statistics affects the form of the cross-section and B has to be (anti)symmetrized into B_θ.

The collision operator (69) has been introduced by Nordheim in 1928 [38] and by Uehling–Uhlenbeck in 1933 on the basis of purely phenomenological considerations [42].

Plugging in the hierarchy (46) an initial datum satisfying (66), we can follow the same procedure as for the Maxwell–Boltzmann statistics, namely we write the full perturbative series expansion expressing $f_j^N(t)$ in terms of the initial datum and try to analyze its asymptotic behavior.

As we did before, we first restrict our attention to those terms of degree less than two in the potential.

The analysis up to second order is performed in [5]. We actually recover here Eq. (68), (69) with the suitable B_θ. Now the number of terms to control is much larger due to the sum over all permutations that enters the definition (66) of the initial state. Also, the asymptotics is much more delicate. In particular, we stress the fact that the initial datum brings its own highly oscillatory factors in the process, contrary to the Maxwell–Boltzmann case where the initial datum is uniformly smooth, and where the oscillatory factors simply come from the collision operators T and C.

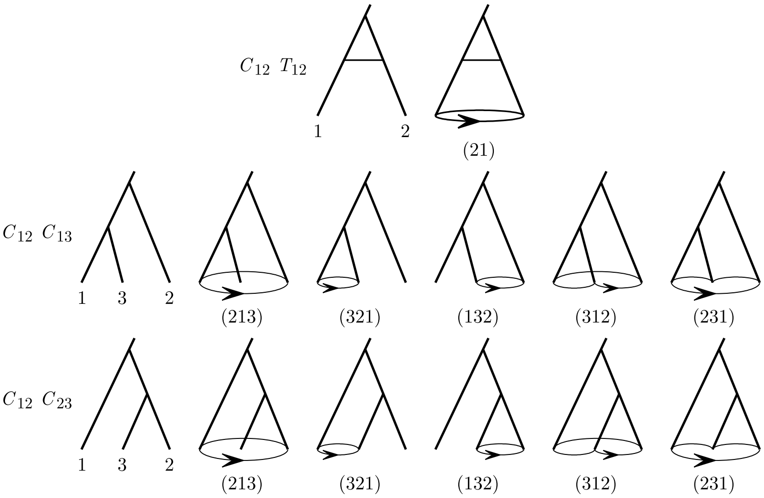
In [5] we consider the graphs of second order in $\hat{\phi}$, show in Fig. 5, which, because of the permutation of initial state, yields various terms: two of them are bilinear in the initial condition f_0 ($C_{12}T_{12}$), and give in the limit the part of Q_θ quadratic in f, twelve are cubic in $f_0(C_{12}C_{13},\ C_{12}C_{23})$. Due to a non-stationary phase argument the two terms with permutation $\pi = id = (123)$ vanish, as that the term with $\pi = (321)$ for $C_{12}C_{13}$, and the term with $\pi = (132)$ for $C_{12}C_{23}$. The two terms with $\pi = (321)$ give rise to truly diverging contributions (negative powers of ε), however their sum is seen to cancel. Last, permutations without fixed point ($\pi = (312), (231)$) and $\pi = (132)$ for $C_{12}C_{13}$, and $\pi = (321)$ for $C_{12}C_{23}$, give the part of Q_θ cubic in f. This ends up the analysis of terms up to second order in the potential.

The computation is heavy and hence we address the reader to [5] for the details.

We mention that a similar second order analysis, using commutator expansions in the framework of the second quantization formalism, has been performed in [27] (following [26]) in the case of the van Hove limit for lattice systems (that is the same as the weak-coupling limit, yet without rescaling the distances). For more recent formal results in this direction, but in the context of the weak-coupling limit, we also quote [21].

We finally observe that the initial value problem for Eq. (68) is somehow trivial for Fermions. Indeed we have the a priori bounds $f \leq 1/(8\pi)^3$ making everything easy. For Bosons the situation is much more involved even for the spatially homogeneous case. The statistics favor large value of f and it is not clear whether the equation can explain dynamical condensation. See, for the mathematical side, [33,34].

Under the weak coupling limit we derived two very different equations, the Fokker–Planck–Landau equation in the classical case, and a Boltzmann equation with a quantum cross section in the quantum case. For the original large particle system a classical or a quantum description



Scaling Limits of Large Systems of Non-linear Partial Differential Equations, Figure 5
The three graphs of second order with the statistical permutations

depends on the value of the physical scales of the macroscopic variables with respect to $\hbar$. The same happen for the two asymptotic equations. The dependence on $\hbar$ of the collision term (69) is given by

$$Q_{UU}(f)(v) = \frac{1}{16\pi^2\hbar^4}\int \mathrm{d}v_1 \int \mathrm{d}\omega \left[\hat{\phi}\left(\tfrac{\omega\,(\omega\cdot w)}{\hbar}\right) + \theta\hat{\phi}\left(\tfrac{w-\omega\,(\omega\cdot w)}{\hbar}\right)\right]^2 \{(1\pm(2\pi\hbar)^3 f)(1\pm(2\pi\hbar)^3 f_1) \cdot f' f_1' - (1\pm(2\pi\hbar)^3 f')(1\pm(2\pi\hbar)^3 f_1') f f_1\} \,. \tag{70}$$

The quantum-cross section term $\hat{\phi}(\omega\,(\omega\cdot(v-v_1))\,/\,\hbar)$ make $\omega\cdot(v-v_1) = O(\hbar)$, so that the collision operator concentrates on grazing collisions. In this sense, the classical limit $\hbar \to 0$ for (70) is a natural grazing collision limit. More formally, the operator (70) tends to the operator Q_L given in Eq. (29) (see [9]); nevertheless no results are available for the limit of the solutions of the equations.

We conclude this section with some consideration on the low-density limit.

In the classical case the low-density limit (or the Boltzmann–Grad limit) yields the usual Boltzmann equation for classical systems and this result has been proved for short times [31]. It is natural to investigate what happens, in the same scaling limit, to a quantum system. Here, due to the fact that the density is vanishing, the particles are too rare to make the statistical correlations effective. As a consequence, we expect that the Bose–Einstein, and Fermi–Dirac statistics, as well fully uncorrelated states, all give rise to the same Boltzmann equation along the low-density limit with the true quantum cross-section, given by the sum of the Born series, under smallness assumption for the potential. The analysis of the partial series of the dominant terms (uniform bounds and convergence as for the weak-coupling limit) has been performed in [6].

Weak-Coupling Limit for a Single Particle: The Linear Theory

Consider the time evolution of a single particle, in the low-density or in the weak-coupling regime, under the action of a random configuration of obstacles $\mathbf{c} = \{c_1 \dots c_N\} \subset \mathbb{R}^{3N}$.

More precisely, after the rescaling, the basic equations are:

$$\dot{x}(t) = v(t)\,, \quad \dot{v}(t) = -\sum_j \nabla\phi_\varepsilon(x(t)-c_j) \tag{71}$$

for a classical particle and, for a quantum particle,

$$i\varepsilon\partial_t\psi = -\frac{\varepsilon^2}{2}\Delta\psi + \sum_j \phi_\varepsilon(x(t)-c_j)\psi\,, \tag{72}$$

where $\phi_\varepsilon = \phi(\frac{x}{\varepsilon})$ and $\phi_\varepsilon = \sqrt{\varepsilon}\phi(\frac{x}{\varepsilon})$ for the low-density and weak-coupling limits respectively. Here ϕ denotes a given smooth potential.

We are interested in the behavior of

$$f_\varepsilon(x,v,t) = \mathbb{E}[f_{\mathbf{c}}(x,v,t)] \tag{73}$$

where $f_{\mathbf{c}}(t)$ is the time evolved classical distribution function or the Wigner transform of ψ according to Eq. (71)

or (72) respectively. Finally $\mathbb{E}$ denotes the expectation with respect to the obstacle distribution, for which, a natural choice could be the Poisson distribution of density μ_ε which is also scaled according to

$$\mu_\varepsilon = \varepsilon^{-2}\,, \quad \mu_\varepsilon = \varepsilon^{-3}$$

for the low-density and weak-coupling respectively.

For the low-density scaling (this is the so called Lorentz model), we obtain, for classical systems, a linear Boltzmann equation (see [3,13,17,23,40]). It is also known that the system does not homogenize to a jump process given by a linear Boltzmann equation in case of a periodic distribution of obstacles [10]. For recent results in this direction see [11].

For the weak-coupling limit of a classical particle we obtain a linear Landau equation as it is shown in [29] and [16].

Let us spend a few words on these results. Each solution $(x(t), y(t)) = (x_c(t), v_c(t))$ of Eq. (71) is a sample of a stochastic process. Looking at the behavior of $v_c(t)$ we suddenly realize that it does not enjoy the Markov property because, once an obstacle produces a change in velocity, we know its existence and this creates time correlations with the future behavior of the particle. However if a recollision with the same obstacle is an unlikely event, we could consider the velocity change due to the collisions, as independent and Poisson distributed (as the obstacles) random variables. In this case it is not difficult to prove the convergence to a diffusion process, solution to the following stochastic differential equation

$$\begin{aligned} dx &= v dt\,, \\ dv &= -2A\frac{v}{|v|^3}dt + \sqrt{\frac{2A}{|v|}}\left(I - \frac{v\otimes v}{|v|^2}\right) dw\,, \end{aligned} \tag{74}$$

where dw is a brownian motion, A is the constant given by Eq. (27), and $\sqrt{\frac{2A}{|v|}}\cdot\left(I - v\otimes v/|v|^2\right)$ is the matrix $\sqrt{2a}$, with a given by Eq. (26). Notice that each collision preserves the energy so that the above process lives on the surface of the sphere $|v| = \text{const}$.

Therefore the main point in the proof of the weak-coupling limit is to show that, in the same limit, the process under consideration is stochastically equivalent to a Markov process with the right properties. Moreover, as a consequence, we have also that $f_\varepsilon(t) \to f(t)$ being f solution to the following linear Landau equation

$$(\partial_t + v\cdot\nabla_x)\, f = \operatorname{div}_v[a(v)\nabla_v f]\,, \tag{75}$$

Alternatively we could proceed differently, in the same (Cauchy–Kovalevskii) spirit of our previous analysis for the nonlinear case. Namely we first represent the time evolved distribution $f_c(t)$ in terms of a series expansion. Then, taking the expectation of each term of the series, we exploit its limit by using the same arguments as in Sect. "Introduction". Finally the convergence of the series (uniformly in ε) should also be proven. This is certainly a possible program which is, however, conceptually and technically weaker than that based on the control of the particle trajectory illustrated above. Indeed the absolute control of the series expansion yielding the solution to Eq. (75) does not use the positivity of the diffusion coefficient so that, at best, we could recover the result for a short time only and in spaces of analytic functions, not at all natural in our context.

This remark applies as well to the nonlinear case for which, however, we do not have any result, even for short times.

As regards the corresponding weak-coupling quantum problem, the easiest case is when ϕ is a Gaussian process. The kinetic equation is still a linear Boltzmann equation. The first result, holding for short times, has been obtained in [41] (see also [30]). More recently this result has been extended to arbitrary times [22]. The technique of [22] can be also applied to deal with a Poisson distribution of obstacles [12] Obviously the cross section appearing in the Boltzmann equation is the one computed in the Born approximation. Finally in [20] the low-density case has been successfully approached. The result is a linear Boltzmann equation with the full cross-section.

Future Directions

The program of obtaining macroscopic equations from the microscopic dynamics is an ambitious and difficult problem. It arised in 1900 with the Hilbert's speech at the congress of mathematicians in Paris, as the sixth problem. After more than hundred years, such a program is far to be achieved.

It is probably true that new ideas and techniques are needed. As we said before in discussing the linear problems, it may be quite possible that classical ideas in the field of partial differential equations are not enough to approach this kind of problems successfully. On the other hand a deeper and more detailed analysis of the particle dynamics is technically difficult and philosophically paradoxical. Indeed from the practical side we introduce a partial differential equation to reduce the difficulties of the study of particle evolution, while its justification may require a deeper control of the underlying microscopic dynamics.

Bibliography

Primary Literature

1. Arseniev AA, Buryak OE (1990) On a connection between the solution of the Boltzmann equation and the solution of the Landau–Fokker–Planck equation (Russian). Mat Sb 181(4):435–446 (translation in Math USSR-Sb 69(2):465–478 1991)
2. Balescu R (1975) Equilibrium and Nonequilibrium Statistical Mechanics. Wiley, New York
3. Boldrighini C, Bunimovich LA, Ya Sinai G (1983) On the Boltzmann equation for nthe Lorentz gas. J Stat Phys 32:477–501
4. Benedetto D, Castella F, Esposito R, Pulvirenti M (2004) Some Considerations on the derivation of the nonlinear Quantum Boltzmann Equation. J Stat Phys 116(114):381–410
5. Benedetto D, Castella F, Esposito R, Pulvirenti M (2005) On The Weak–Coupling Limit for Bosons and Fermions. Math Mod Meth Appl Sci 15(12):1–33
6. Benedetto D, Castella F, Esposito R, Pulvirenti M (2006) Some Considerations on the derivation of the nonlinear Quantum Boltzmann Equation II: the low-density regime. J Stat Phys 124(2–4):951–996
7. Benedetto D, Castella F, Esposito R, Pulvirenti M (2008) From the N-body Schrödinger equation to the quantum Boltzmann equation: a term-by-term convergence result in the weak coupling regime. Commun Math Phys 277(1):1–44
8. Benedetto D, Esposito R, Pulvirenti M (2004) Asymptotic analysis of quantum scattering under mesoscopic scaling. Asymptot Anal 40(2):163–187
9. Benedetto D, Pulvirenti M (2007) The classical limit for the Uehling–Uhlenbeck operator. Bull Inst Math Acad Sinica 2(4):907–920
10. Burgain J, Golse F, Wennberg B (1998) On the distribution of free path lenght for the periodic Lorentz gas. Comm Math Phys 190:491–508
11. Caglioti E, Golse F (2003) On the distribution of free path lengths for the periodic Lorentz gas. III Comm Math Phys 236(2):199–221
12. Chen T (2005) Localization lengths and Boltzmann limit for the Anderson model at small disorders in dimension 3. J Stat Phys 120(1–2):279–337
13. Caglioti E, Pulvirenti M, Ricci V (2000) Derivation of a linear Boltzmann equation for a periodic Lorentz gas. Mark Proc Rel Fields 3:265–285
14. Cercignani C, Illner R, Pulvirenti M (1994) The mathematical theory of dilute gases, Applied Mathematical Sciences, vol 106. Springer, New York
15. Degond P, Lucquin–Desreux B (1992) The Fokker–Planck asymptotics of the Boltzmann collision operator in the Coulomb case. Math Models Methods Appl Sci 2(2):167–182
16. Dürr D, Goldstain S, Lebowitz JL (1987) Asymptotic motion of a classical particle in a random potential in two dimension: Landau model. Comm Math Phys 113:209–230
17. Desvillettes L, Pulvirenti M (1999) The linear Boltzmann eqaution for long-range forces: a derivation for nparticle systems. Math Moduls Methods Appl Sci 9:1123–1145
18. Di Perna RJ, Lions PL (1989) On the Cauchy problem for the Boltzmann equatioin. Ann Math 130:321–366
19. Esposito R, Pulvirenti M (2004) From particles to fluids. In: Friedlander S, Serre D (eds) Handbook of Mathematical Fluid Dynamics, vol 3. Elsevier, North Holland, pp 1–83
20. Eng D, Erdös L (2005) The linear Boltzmann equation as the low-density limit of a random Schrödinger equation. Rev Math Phys 17(6):669–743
21. Erdös L, Salmhofer M, Yau HT (2004) On the quantum Boltzmann equation. J Stat Phys 116:367–380
22. Erdös L, Yau HT (2000) Linear Boltzmann equation as a weak-coupling limit of a random Schrödinger equation. Comm Pure Appl Math 53:667–735
23. Gallavotti G (1972) Rigorous theory of the Boltzmann equation in the Lorentz gas in Meccanica Statistica. reprint Quaderni CNR 50:191–204
24. Goudon T (1997) On Boltzmann equations and Fokker–Planck asymptotics: influence of grazing collisions. J Statist Phys 89(3–4):751–776
25. Grad H (1949) On the kinetic Theory of rarefied gases. Comm Pure Appl Math 2:331–407
26. Hugenholtz MN (1983) Derivation of the Boltzmann equation for a Fermi gas. J Stat Phys 32:231–254
27. Ho NT, Landau LJ (1997) Fermi gas in a lattice in the van Hove limit. J Stat Phys 87:821–845
28. Illner R, Pulvirenti M (1986) Global Validity of the Boltzmann equation for a two-dimensional rare gas in the vacuum. Comm Math Phys 105:189–203 (Erratum and improved result, Comm Math Phys 121:143–146)
29. Kesten H, Papanicolaou G (1981) A limit theorem for stochastic acceleration. Comm Math Phys 78:19–31
30. Landau LJ (1994) Observation of quantum particles on a large space-time scale. J Stat Phys 77:259–309
31. Lanford III O (1975) Time evolution of large classical systems. In: Moser EJ (ed) Lecture Notes in Physics 38. Springer, pp 1–111
32. Lifshitz EM, Pitaevskii LP (1981) Course of theoretical physics "Landau–Lifshit", vol 10. Pergamon Press, Oxford-Elmsford
33. Xuguang LU (2005) The Boltzmann equation for Bose–Einstein particles: velocity concentration and convergence to equilibrium. J Stat Phys 119(5–6):1027–1067
34. Xuguang LU (2004) On isotropic distributional solutions to the Boltzmann equation for Bose–Einstein particles. J Stat Phys 116(5–6):1597–1649
35. CB Jr Morrey (1955) On the derivation of the equations of hydrodynamics from statistical mechanics. Comm Pure Appl Math 8:279–326
36. Nier F (1996) A semi-classical picture of quantum scattering. Ann Sci Ec Norm Sup 29(4):149–183
37. Nier F (1995) Asymptotic analysis of a scaled Wigner equation and quantum scattering. Transp Theory Statist Phys 24(4–5):591–628
38. Nordheim LW (1928) On the Kinetic Method in the New Statistics and Its Application in the Electron Theory of Conductivity. Proc Royal Soc Lond Ser A 119(783):689–698
39. Spohn H (1991) Large scale dynamics of interacting particles, Texts and monographs in physics. Springer
40. Spohn H (1978) The Lorentz flight process converges to a random flight process. Comm Math Phys 60:277–290
41. Spohn H (1977) Derivation of the transport equation for electrons moving through random impurities. J Stat Phys 17:385–412
42. Uehling EA, Uhlembeck GE (1933) Transport Phenomena in Einstein–Bose and Fermi–Dirac Gases. Phys Rev 43:552–561
43. Villani C (2002) A review of mathematical topics in collisional kinetic theory. In: Friedlander S, Serre D (eds) Handbook of

Mathematical Fluid Dynamics, vol 1. Elsevier, North Holland, pp 71–307

44. Villani C (1998) On a new class of weak solutions to the spatially homogeneous Boltzmann and Landau equations. Arch Rational Mech Anal 143(3):273–307
45. Wigner E (1932) On the quantum correction for the thermodynamical equilibrium. Phys Rev 40:742–759

Books and Reviews

Balescu R (1975) Equilibrium and Nonequilibrium Statistical Mechanics. Wiley, New York

Cercignani C, Boltzmann L (1998) The man who trusted atoms. Oxford University Press, Oxford

Cercignani C, Illner R, Pulvirenti M (1994) The mathematical theory of dilute gases. Applied Mathematical Sciences, vol 106. Springer, New York

Esposito R, Pulvirenti M (2004) From particles to fluids. In: Friedlander S, Serre D (eds) Handbook of Mathematical Fluid Dynamics, vol 3. Elsevier, North Holland, pp 1–83

Spohn H (1991) Large scale dynamics of interacting particles, Texts and monographs in physics. Springer, Heidelberg

Villani C (2002) A review of mathematical topics in collisional kinetic theory. In: Friedlander S, Serre D (eds) Handbook of Mathematical Fluid Dynamics, vol 1. Elsevier, North Holland, pp 71–307

Scaling Properties, Fractals, and the Renormalization Group Approach to Percolation

DIETRICH STAUFFER
Institute for Theoretical Physics, Cologne University,
Köln, Germany

Article Outline

Glossary

Cluster Clusters are sets of occupied neighboring sites.

Critical exponent At a critical point or second-order phase transition, many quantities diverge or vanish with a power law of the distance from this critical point; the critical exponent is the exponent for this power law.

Fractals Fractals have a mass varying with some power of their linear dimension. The exponent of this power law is called the fractal dimension and is smaller than the dimension of the space.

Percolation Each site of a large lattice is randomly occupied or empty.

Renormalization A cell of several sites, atoms, or spins is approximated by one single site etc. At the critical point, these supersites behave like the original sites, and the critical point thus is a fixed point of the renormalization.

Definition of the Subject

Percolation theory mostly deals with large lattices where every site is randomly either occupied or empty. In particular it studies the resulting clusters which are sets of neighboring occupied sites.

Introduction

Paul Flory, who later got the Chemistry Nobel prize, published in 1941 the first percolation theory [1], to describe the vulcanization of rubber [2]. Others later applied and generalized it, in particular by dealing with percolation theory on lattices and by studying it with computers. Most of the theory presented here was known around 1980, though in the case of computer simulation with less accuracy than today. But on the questions of universality, of critical spanning probability and of the uniqueness of infinite clusters, the 1990's have shown some of our earlier opinions to be wrong. And even today it is questioned by some that the critical exponents of percolation theory can be applied to real polymer gelation, the application which Flory had in mind two-thirds of a century ago.

On a large lattice we assume that each site independently and randomly is occupied with probability p and empty with probability $1 - p$. Depending on applications, also other words can be used instead of occupied and empty, e. g. Republican and Democrat for the majority party in an electoral district of the USA. A *cluster* is now defined as set of occupied neighboring sites. Percolation theory deals with the number and structure of these clusters, as a function of their size s, i. e. of the number s of occupied sites in the cluster. In particular it asks whether an infinite cluster spans from one side of the lattice to the opposite side. Alternatively, and more naturally if one wants to describe chemical reactions for rubber vulcanization, this site percolation can be replaced by bond percolation, where every site is occupied but the link between neighboring sites is either present with probability p or absent with probability $1 - p$, again independently and ran-

Scaling Properties, Fractals, and the Renormalization Group Approach to Percolation, Table 1
Site and bond percolation thresholds for one dimension, three two-dimensional, four three-dimensional and four hypercubic lattices in higher dimensions [1,3]

p_c	Site	Bond
$d = 1$ chain	1	1
honeycomb	.697043	$1 - 2\sin(\pi/18)$
square	.592746	1/2
triangular	1/2	$2\sin(\pi/18)$
diamond	.4301	.3893
SC	.311608	.248813
BCC	.245691	.180287
FCC	.199236	.120163
$d = 4$ hypercubic	.196885	.160131
$d = 5$ hypercubic	.140797	.118172
$d = 6$ hypercubic	.109018	.094202
$d = 7$ hypercubic	.088951	.078675

domly for each link. A cluster is now a set of neighboring sites connected by links, and the size s of the cluster can be counted as the number of links, or as the number of sites, in that cluster. Because of this ambiguity we discuss here mainly site percolation; bond percolation is similar in the sense that it belongs to the same universality class (same critical exponents). One may also combine both choices and study site-bond percolation where each site is randomly occupied or empty, and where each bond between neighboring occupied sites is randomly present or absent.

Neither temperature nor quantum effects enter this standard percolation model, which is purely geometrical probability theory. However, to understand why percolation works the way it does it is helpful to understand thermal phase transitions like the vapor-liquid critical point; and for magnetic applications it is useful to know that some spins (atomic magnetic moments) have only two states, up or down, according to quantum mechanics. We will explain these physics aspects later.

For small p, most of the occupied sites are isolated $s = 1$, coexisting with only few pairs $s = 2$ and triplets $s = 3$. For large p, most of the occupied sites form one "infinite" cluster spanning the lattice from left to right, with a few small isolated holes in it. Thus there exists one percolation threshold p_c such that for $p < p_c$ we have no spanning cluster and for $p > p_c$ we have (at least) one spanning cluster. Inspite of decades of research in this seemingly simple problem, no exact solution for p_c is proven or guessed for site percolation on the square lattice with nearest-neighbor bonds; only numerically we know it to be about 0.5927462. For site percolation on the triangular lattice or bond percolation on the square lattice, $p_c = 1/2$ exactly. More thresholds are given in Table 1 [1]. They are valid in the limit of $L \to \infty$ for lattices with about L^d sites in d dimensions. For small L instead of a sharp transition at p_c one has a rounded changeover: with a very low probability one chain of L occupied sites at $p = 1/L^{d-1}$ spans from left to right. In one dimension, a small chain can easily be spanned if p is close to one, but for $L \to \infty$ the threshold approaches $p_c = 1$ since at smaller p a hole will appear about every $1/(1 - p)$ sites and prevent any cluster to span.

Methods

This section summarizes some of the methods employed to find percolation properties, first by pencil and paper, and then with the help of computers for which Fortran programs are published e. g. in [4,5]. More details on simulations are reviewed by Ziff in this percolation part of this encyclopedia.

Mean Field Limit

The Bethe lattice or Cayley tree neglects all cyclic links and allows a solution with paper and pencil. We start from one central site, and let z bonds emanate from that. At the end of each bond sits a neighbor. Then from each of these neighbors again z bonds emanate, one back to the central site and $z - 1$ to new sites further outward. They in turn lead again each to $z - 1$ new sites, and so on. None of the newly added sites agrees with one of the already existing sites, and so we can travel along the bonds only outwards or back, but never in a loop. It is quite plausible that an infinite cluster of bond percolation is formed if each site leads to at least one more outward site along an existing bond, that means if $(z - 1)p > 1$. This condition also holds for site percolation. Thus

$$p_c = 1/(z - 1) . \qquad (1)$$

In this way Flory calculated the threshold and other percolation properties. Today we call this the "mean field" universality class in analogy with thermal phase transitions. The critical exponents, to be discussed below, are integers or simple fractions. To this universality class belong also the Erdös-Rényi random graphs, where we connect in an assembly of N points each pair with a low probability $\propto 1/N$. And the same universality class is reached if we let the dimension d of the hypercubic lattice go to infinity (or at least take it above 6). A disadvantage of the Bethe lattice is its lack of realism: If the length of the bonds is constant, then the exponential increase of the number of sites and bonds with increasing radius leads to an infinite density.

Small Clusters

The probability of a site to be an isolated $s = 1$ cluster on the square lattice is $n_1 = p(1-p)^4$ since the site must be occupied and all its four neighbors be empty. The formula for pairs is $n_2 = 2p^2(1-p)^6$ since the pair can be oriented horizontally or vertically, resulting in the factor 2. Similar, only more difficult, is the evaluation of n_s with a maximum s usually 10 to 20; the general formula is

$$n_s = \sum_t g_{st}\, p^s\, (1-p)^t \tag{2}$$

where the perimeter t is the number of empty neighbors and g_{st} is the number of configurations (or lattice animals, or polyominoes) of size s and perimeter t. The King's College group in London published these results decades ago. With techniques borrowed from series expansions near thermal critical phenomena, these polynomials allow to estimate not only p_c but also many other quantities (see below) diverging or vanishing near p_c.

Leath Cluster Growth

In the cluster growth method of Leath (1976) one starts with one occupied site in the center of the lattice. Then a cluster is grown by letting each empty neighbor of an already occupied cluster site decide once and for all, whether it is occupied or empty. One needs to keep and to update a perimeter list of undecided neighbors. If that list becomes empty, the cluster growth is finished, and no boundary effects of the lattice influence this cluster. If, on the other hand, the cluster reaches the lattice boundary, one has to stop the simulation and can regard this cluster as spanning (from the center to one of the sides). Repeating many times this growth simulation one can estimate p_c as well as the cluster numbers. More precisely, the cluster statistics obtained in this way is not n_s but $n_s s$ since the original center site belongs with higher probability to a larger than to a smaller cluster.

Hoshen-Kopelman Labelling

To go regularly through a large lattice, which may even be an experimentally observed structure to be analyzed by computer, one could number consecutively each seemingly new cluster, and if no clusters merge later then one has a clear classification: All sites belonging to the first cluster have label 1, all sites of the second cluster have label 2, etc. Unfortunately, this does not work. In the later analysis it may turn out that two clusters which at first seemed separate actually merge and form one cluster:

```
* *     * *         1 1     2 2
  *   * * *           1   3 ? 2
* * * * *           4 ? ? x ?
```

Already in the simple structure shown on the left we have several such label conflicts. The labels to the right come from going through the lattice like a typewriter, from left to right, and after each line to the lower line. When we come to the right neighbor of the 3 we see that 3 is really part of the cluster with label 2. And at the right neighbor of 4 we see that 4 belongs to cluster 1. The stupid method is to go back and to relabel all 3 into 2, and all 4 into 1. If then we come to the site marked with x we see that the whole structure is really one single cluster, and thus all labels 2 have to be relabeled into a 1. This is inefficient for large lattices. Instead, Hoshen and Kopelman (1976) gave each site label $m = 1, 2, 3, \ldots$ another index $n(m)$. This label $n(m)$ of labels equals its argument, $n(m) = m$, if it is still a good "root label", and it equals another number k is the cluster with initial label m later turned out to be part of an earlier cluster k. By iterating the command `m = n(m)` until finally the new m equals $n(m)$ one finds this root label. For the above we make the following assignments and re-assignments to n: $n(1) = 1$, $n(2) = 2$, $n(3) = 3$, $n(3) = 2$, $n(4) = 4$, $n(4) = 1$, $n(2) = 1$. Clusters are now characterized by the same root label for all their labels.

An advantage if this method is that only one line of the square lattice, or one hyperplane of the d-dimensional lattice, needs to be stored at any time, besides the array $n(m)$. And that array can also be reduced in size by regular recycling no longer used labels n, just as beer bottles can be recycled. Lattices with more than 10^{13} sites were simulated, using parallel computers. However, understanding the details of the algorithms and finding errors in them can be very frustrating.

Sometimes one wants to determine the cluster numbers for numerous different p from 0 to 1. Instead of starting a new analysis for each different p one may also fill the lattice with new sites, and make the proper labeling of labels whenever a new site was added [6]. Similarly, one can determine the properties of various lattice sizes L by letting L grow one by one and relabeling the cluster after each growth step [7]. Unfortunately, these two methods came long after most of the percolation properties were already studied quite well by standard Hoshen-Kopelman analysis.

Relation to Ising and Potts Models

The relation between percolation and thermal physics was useful for both sides: Scaling theories for percolation could follow scaling theories for thermal physics from ten years earlier, and computer simulations for thermal physics could use the Leath and Hoshen-Kopelman algorithms of cluster analysis, leading to the Wolff and Swendsen-Wang methods, respectively, a decade later. A mathematical foundation is given by the Kasteleyn-Fortuin theorem [8] for the partition function Z of the Q-state Potts model at temperature T:

$$Z(Q) = \langle Q^N \rangle \tag{3}$$

where N is the total number $\sum_s n_s$ of clusters for bond percolation at probability $1 = \exp(-2J/k_B T)$, $\langle \ldots \rangle$ indicates an average over the configurations at this probability, k_B is Boltzmann's constant and 2J is the energy needed to break a bond between neighboring spins. (Each site i of a Potts lattice carries a variable $S_i = 1, 2, \ldots Q$; the energy of a neighbor pair is $-2J$ if the two variables agree, and zero otherwise.) For the special Ising case $Q = 2$ see also this author in this encyclopedia, "Opinion dynamics..." and "Phase transitions...".

Q values of 3 and larger are interesting since for increasing Q a second-order phase transition with a continuous order parameter changes into a first-order phase transition with a jumping order parameter, when T increases. The special case $Q = 2$ is the spin 1/2 Ising model (the model is pronounced EEsing, not EYEsing since Ernst Ising was born in Cologne, Germany, and became US citizen Ernest Ising only after publishing his theory in 1925 and surviving Nazi persecution 1933–1944). The limit $Q \to 0$ corresponds to some tree structures (no cyclic links, as in Flory's percolation theory, [9]). Percolation, on the other hand, corresponds to the limit $Q \to 1$, in the following way: The "free energy" in units of $k_B T$ is in this limit $\ln Z = \ln\langle \exp(N \ln Q)\rangle \simeq \ln\langle \exp[(Q-1)N]\rangle \simeq \ln\langle 1 + (Q-1)N\rangle \simeq (Q-1)N$. Thus for Q near unity this thermal free energy, divided by $Q - 1$, is the number of percolation clusters.

In this way thermal physics and percolation are related, and the cluster numbers N correspond to a free energy. In thermal physics, the negative derivative of the free energy with respect a conjugate field gives the order parameter (e.g magnetic field and magnetization), and the field derivative of the order parameter is called the susceptibility. For liquid-gas equilibria, the order parameter is the volume (or the density), the field is the pressure (or chemical potential), and the analog of the susceptibility is the compressibility. This result Eq. (3), not its derivation, we should keep in mind if we now look at the percolation quantities of interest.

Formally we may define for percolation a free energy F as a generating function of a ghost field h:

$$F(h) = \sum_s n_s \exp(-hs) \; . \tag{4}$$

Then its first h-derivative is $-\sum_s n_s s$, and the second one is $\sum_s n_s s^2$, sums which appear below in the percolation probability P_∞ (the order parameter) and the mean cluster size $S = \sum_s n_s s^2 / \sum_s n_s s$ (the susceptibility).

Quantities and Exponents

The basic quantity is n_s, the number (per site) of clusters containing s sites each, and often is an average over several realizations for the same occupation probability p in the same lattice. Several moments

$$M_k = \sum_s n_s s^k \tag{5}$$

are used to define other quantities of interest; in these sums the infinite (spanning) clusters are omitted. The following proportionalities are valid asymptotically in the limit of large lattice size L and for $p \to p_c$:

$$F = M_0 \propto |p - p_c|^{2-\alpha} + \ldots \tag{6a}$$

$$P_\infty = p - M_1 \propto (p - p_c)^\beta \tag{6b}$$

$$S = M_2/M_1 \propto |p - p_c|^{-\gamma} \; . \tag{6c}$$

Here F is the analog of the thermal free energy, where the three dots represent analytic background terms whose derivatives are all finite. Since every occupied site must belong either to a finite or to an infinite cluster, $P_\infty = p - \sum_s n_s s$ is the fraction of sites belonging to the infinite cluster and gives the probability that from a randomly selected site we can walk to a lattice boundary along a path of occupied sites. It is thus called the percolation probability but needs to be distinguished from the probability p that a single site is occupied and from the probability R, with $R(p < p_c) = 0, \;\; R(p > p_c) = 1$, that there is a spanning cluster in the lattice.

The quantity S is usually called the mean cluster size, and we follow this tradition even though it is very bad. There are many ways to define a mean size, and polymer chemists have the much more precise notation of a number average M_1/M_0, a weight average M_2/M_1 and a z average M_3/M_2 for the cluster size (= degree of polymerization). Physicists arbitrarily call the weight-averaged s the mean cluster size S. Numerically, the exponent γ is determined more easily from the "susceptibility" $\chi = M_2 \propto |p - p_c|^{-\gamma}$, since the denominator M_1 in Eq. (6c) approaches very slowly its asymptotic limit of 1.

Scaling Properties, Fractals, and the Renormalization Group Approach to Percolation, Table 2

Critical exponents for percolation clusters. The mean field values are valid for six and more dimensions and also apply to Flory's Bethe approximation and to Erdös-Rényi random graphs. The exponents α, δ, σ, τ can be derived from the scaling laws, Eq. (8)

d	β	γ	ν
2	5/36	43/18	4/3
3	0.41	1.796	0.88
≥ 6	1	1	1/2

The radius of a cluster R_s can be defined as the rms distance $r_i,\ i = 1, 2, \ldots, s$ of cluster sites from the center of mass r_c of the cluster (radius of gyration):

$$R_s^2 = \langle \sum_i (r_i - r_c)^2/s \rangle \tag{6d}$$

where the $\langle \ldots \rangle$ average over all cluster configurations at probability p. Then the correlation length ξ is related to the z-average cluster radius through

$$\xi^2 = \sum_s R_s^2 n_s s^2 / \sum_s n_s s^2 \propto |p - p_c|^{-\nu} \tag{6e}$$

with another critical exponent ν.

Finally, right at $p = p_c$, the cluster numbers decay as

$$n_s \propto 1/s^{2+1/\delta} \tag{7}$$

where δ must be positive to allow a finite density $\sum_s n_s s = p$.

These five critical exponents are not independent of each other but are related in d dimensions through the scaling laws:

$$2 - \alpha = \gamma + 2\beta = (\delta + 1)\beta = d\nu \tag{8a}$$

as known from thermal phase transitions; the last equation involving d is not valid in mean field theory (large d) but only for $d \leq 6$. Table 2 gives the numerical estimates of the exponents in three dimensions as well as their mean field values for $d \geq 6$ and their exact two-dimensional results [10,11]. Thus, for six and less dimensions, if you know two exponents you know them all; thus far.

These scaling laws (8a) can be derived by assuming

$$n_s = s^{-\tau} f[(p - p_c)s^{\sigma}] \quad (\tau = 2 + 1/\delta,\ 1/\sigma = \beta\delta) \tag{8b}$$

which was first postulated for the thermal Ising model, and then successfully applied to percolation. Here f is a suitable scaling function, which only in the mean-field limit approaches a Gaussian.

For both thermal critical phenomena and percolation, "universality" asserts that these critical exponents are independent of many details and (for the Potts model) depend only on the dimensionality d and the number Q of possible spin states. Since percolation corresponds to $Q \to 1$ this means that the exponents depend only on d. There are exceptions from this universality for thermal phase transitions, but for random percolation thus far it worked. However, the numerical value of the percolation threshold p_c is not a critical exponent, depends on the lattice structure, and is different for site and bond percolation.

This universality is one of the reasons why the investigation of exponents is important: They allow to classify models and materials. Similarly, in biology we have many birds of different colors, and many types of domestic animals. Biology became a systematic science only when it was found that all mammals share certain properties, which birds no not have. Thus there is the universality class of mammals.

(The proportionality factors in Eq. (6) are not universal, but some of their combinations are; for example, the ratio of the proportionality factors for S above to below p_c is universal. In some sense also the probability $R(p = p_c)$ of a lattice to contain one spanning cluster at the threshold is universal: same for bond and site percolation; however, that probability depends on the boundary conditions and the shape of the sample and thus is far less universal that the mentioned ratio for S.)

Unfortunately, there is another exponent which does not follow from the cluster numbers and radii and for which no scaling law is accepted which relates it to the other exponents above. This refers to the electrical conductivity

$$\Sigma \propto (p - p_c)^{\mu} \tag{9}$$

when each occupied site (or bond) conducts electrical current and each empty site (or deleted bond) is an insulator. The numerical values are 1.30, 2.0 and 3 in two, three and at least six dimensions. If bonds are realized by elastic springs with bending forces, the elastic exponent may be $\mu + 2\nu$ if entropy effects are negligible, or $2 - \alpha$ if entropy effects are dominant. Moreover, μ is less universal: the above lattice values do not hold on a continuum (conducting spheres which may overlap). Similarly, the kinetics of the Ising model determine a critical exponent which differs in different variants of the kinetics and may not be related to the static Ising exponents like β and γ.

Fractal Dimension; Incipient Infinite Cluster

Fractal Dimension D

Typical objects of geometry classes in school are one-dimensional lines, two-dimensional squares or circles, and three-dimensional cubes or spheres. They have a length (radius) L and a mass (volume for unit density) M with $M \propto L^d$ for d dimensions. In reality, mother nature produces much more complex objects, like trees, where the mass varies with a power of the tree height below 3:

$$M \propto L^D \quad (D < d,\ L \to \infty)\,. \tag{10a}$$

D is the fractal dimension, and such objects are called fractals, particularly if they also are self-similar in that a small twig looks like a big branch, etc. Finite-size scaling theory then relates D of the largest (spanning?) cluster at $p = p_c$ to the above percolation exponents through

$$D = d - \beta/\nu = (\gamma + \beta)/\nu = 1/(\sigma\nu) = d/(1 + 1/\delta) \tag{10b}$$

for $d \leq 6$. Thus the critical cluster is about 1.9-dimensional in two and 2.5-dimensional in three dimensions, while in the mean field regime for $d \geq 6$ we have $D = 4$. Why is this so? Any quantity X which is supposed to vary near $p = p_c$ as $|p - p_c|^x$ does so only for infinitely large systems. For a finite lattice size L, the transition is rounded, and neither X nor any of its p-derivatives diverges or becomes exactly zero. In particular, the typical cluster radius or correlation length $\xi \propto |p - p_c|^{-\nu}$ cannot become infinite but becomes of order L. Then the relation $X \propto \xi^{-x/\nu}$ is replaced by

$$X(p = p_c) \propto L^{-x/\nu} \tag{11a}$$

at the threshold, and

$$X(p \simeq p_c) = L^{-x/\nu} g[(p - p_c)L^{1/\nu}] \tag{11b}$$

near the threshold, with a suitable scaling function g. In particular, the fraction P_∞ of sites belonging to the largest cluster at $p = p_c$ vanishes as $L^{-\beta/\nu}$, and the total number M of sites in this cluster as

$$M \propto L^{d-\beta/\nu} \quad \text{or} \quad D = d - \beta/\nu \tag{11c}$$

as asserted in Eq. (10b).

Figure 1 shows the second moment $\chi = M_2 = \sum_s n_s s^2$ in small (curve) and large (+) simple cubic lattices, differing only for $p \simeq p_c$. Figure 2 shows right at $p = p_c$ the variation with lattice size of the number M of sites in the largest cluster and of the second moment M_2 (susceptibility).

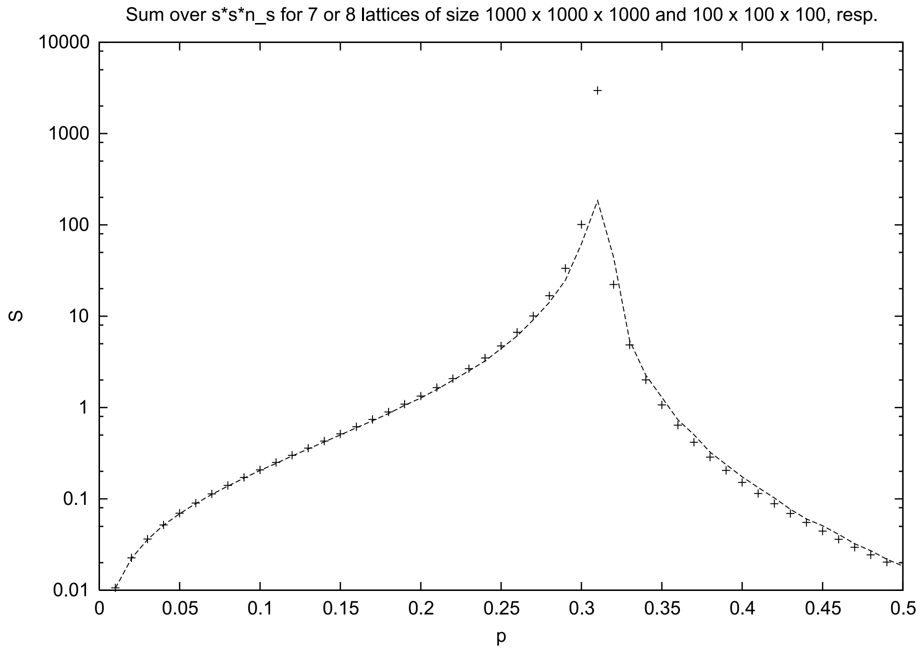
In a finite lattice, the probability $R(p)$ of a spanning cluster to exist goes from nearly zero to nearly unity in a p-interval proportional to $1/L^{1/\nu}$, according to Eq. (11b) with $x = 0$. The derivative $\mathrm{d}R/\mathrm{d}p$ is the probability that spanning first occurred at probability p. It is plausible that this probability, peaked around p_c, is a Gaussian, i. e. a normal distribution. Unfortunately, the Evil Empire, also known as the Departments of Chemical Engineering, destroyed [12] this beautiful idea: Since for $p \simeq p_c$ and $\xi \sim L$ every part of the lattice is correlated with the rest of the lattice, the central limit theorem does not hold.

(If for $p \ll p_c$ we let the cluster size s go to infinity, which requires a special algorithm, we get into the universality class of lattice animals, Sect. "Small Clusters". Most simply, in the limit $p \to 0$, Eq. (2) simplifies to $n_s/p^s = g_{st}$, that means we look at the distribution of configurations with s sites and perimeter t, where all configurations of a given s are weighted equally, whatever their perimeter t is. An important result for these animals is that in three dimensions their radius R_s varies as $\sqrt{s}$, i,e. their fractal dimension is exactly 2. In two dimensions, only numerical estimates exist with $D \simeq 1.56$. It is highly unusual that a problem has an exact solution in three but not in two dimensions.)

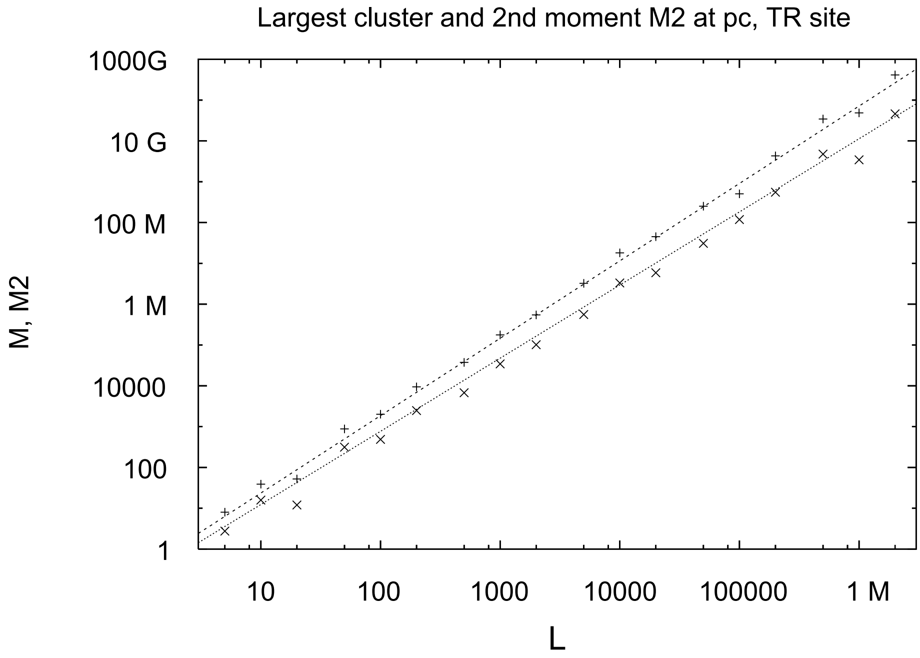
Incipient Infinite Cluster

Right at $p = p_c$ the largest cluster spans the lattice with a pseudo-universal probability $0 < R(p_c) < 1$, and then has a density P_∞ going to zero for L going to infinity. It is also called the incipient infinite cluster IIC. Most of the IIC consists of dangling ends which carry no current if the cluster is interpreted as a random resistor network with conductivity Σ, see Eq. (9) above. The remaining current carrying "backbone" has a fractal dimension 1.643 in two dimensions, 1.7 in three and 2 in at least six dimensions and mostly consists of blobs where current flows along several parallel though connected paths. The few "articulation" sites or bonds, the removal of which cuts the network into two or more parts, are also called "red" since all the current flows through them; they have a fractal dimension of only $1/\nu = 0.75$, 1.14 and 2 in two, three and $\geq$ six dimensions.

How many infinite clusters do we have? The easy answer is: none below, perhaps one at and always one above p_c in an infinite network. Indeed, this is what was claimed mathematically in the 1980's [13]: The number of infinite clusters is zero, one or infinite. Later mathematics excluded the last choice of infinitely many clusters, even though in seven dimensions scaling arguments, confirmed



Scaling Properties, Fractals, and the Renormalization Group Approach to Percolation, Figure 1
"Susceptibility" M_2 in simple-cubic lattice. For the smaller size the maximum is reduced appreciably



Scaling Properties, Fractals, and the Renormalization Group Approach to Percolation, Figure 2
Number M of sites in largest cluster (+) and susceptibility M_2 (x) at $p = p_c = 1/2$ for triangular site percolation. The two straight lines have the exact slopes $D = 91/48$ and $\gamma/\nu = 43/24$ predicted by finite-size scaling. The largest lattice took about 36 h on a workstation with 2 Gb memory. Tiggemann [7] simulated $L = 7 \times 10^6$, 25024, 1305, 225 for $d = 2, 3, 4, 5$ on a large parallel computer

by numerical studies [14], indicated the number of IIC to go to infinity for increasing L in seven dimensions. Only in 1995 and later Aizenman [15] predicted that in all dimensions one may have several spanning clusters at $p = p_c$, in agreement with simulations [16].

Why were the earlier uniqueness theorems irreproducible at p_c and for very elongated rectangles even above p_c [17]? A clear definition of "infinite" is missing in some of the mathematics, although [13] defined a cluster as infinite if its cardinality (= number of sites in it) is infinite for $L \to \infty$ in a hypercubic lattice of L^d sites. Clear definitions of infinity are, of course, needed for reliable proofs [18]. Measure theory as applied in some theorems may be based on some axioms which are not applicable for a fractal IIC. Very simply, imagine each line of an $L \times L$ square lattice to have one randomly selected site occupied and all others empty. The set of occupied site then has cardinality L which is infinite for infinite lattices, but its density becomes zero. Does your measure theory agree with this? More relevant for percolation, even for $p < p_c$ the largest cluster has a size increasing logarithmically with lattice size and thus can be described as infinite, invalidating the percolation threshold as the onset of infinite clusters. Thus infinite might be defined as increasing with a positive power of L, i. e. having a positive fractal dimension. Then we have infinitely many infinite clusters only at $p = p_c$, though in most cases only the largest of them is a spanning cluster. Using "spanning" as a definition of an infinite cluster seems to cause the smallest problems.

Thus one should not regard a question as settled if some mathematical theorem claims to have answered it. The mathematics may not apply to the same problem one is interested in, or (see bootstrap percolation in this encyclopedia) may apply only for unrealistically large lattices. On the other hand, also computer simulations should be relied upon only if confirmed independently. And in the interpretation of simulation results one should be objective and not try to agree with prevailing theories. For example, [14] might already have seen the multiplicity of infinite clusters in five dimensions, not only in seven, had she not followed her obviously incompetent postdoctoral mentor.

(On a more positive side, mathematicians [19] solved biased diffusion on percolating clusters above p_c only a few years after physicists still had controversies about their simulations.)

Simple Renormalization Group

Why are scaling laws and finite-size scaling so simple? Why is universality valid for the exponents? These question arose for thermal critical phenomena as well as for percolation. The main reason is that the correlation length ξ goes to infinity at the critical point. Thus all approximations which restrict the correlations to some finite lengths eventually become wrong, and instead the scaling ideas become correct. They were explained by Ken Wilson through what he called renormalization group, around 1970, and he got the physics Nobel prize for it in 1982. Basically, since correlations extend over long distances, the single atom or lattice point becomes irrelevant and can be averaged over. In politics, we have a similar effect: Many democracies are based on electoral districts, and the candidate winning most votes within this district represents this district in the national parliament. It is the cooperation of many people within the electoral district, not the single vote, which is decisive.

Returning to an $L \times L$ lattice, we can divide it into many blocks of linear dimension b, and treat a block analogously to an electoral district. Thus in an Ising model, if the majority of block spins point upward, the whole block is represented by a superspin pointing up, analogous to the single representative in politics. These block spins then act like the original spins, one can put $b \times b$ superspins into one superblock, and have just one super-representative following the majority opinion of the representatives within the superblock. This process can be continued: at each stage $b \times b$ lower representatives are normalized into a single higher representative.

Such a renormalization by majority rule works fine with Ising spins, but percolation deals with connections, not with up and down spins. Thus for percolation a $b \times b$ block is normalized into an occupied supersite if and only if there is a spanning cluster within the block; otherwise the superblock is defined empty. In this way, whole blocks are normalized into single sites via connectedness. And the renormalization is reduced to the standard question which was asked already before Wilson's invention: Does a $b \times b$ lattice have a spanning cluster? The supersite is thus occupied if and only if the block spans, which happens with probability $R_b(p)$. If we call p' the probability of the supersite to be occupied, we thus have

$$p' = R_b(p) \,. \tag{12a}$$

If we are at $p = p_c$, then the renormalization should not change anything drastic since ξ is larger than any b; thus if the renormalization would be exact we would have

$$p'_c = R_b(p_c) \,. \tag{12b}$$

Practically we determine a fixed point $p = p^*$ such that

$$p^* = R_b(p^*) \,. \tag{12c}$$

and then find p_c as the limit of p^* for $b \to \infty$, which again is similar to what percolation experts did before this renormalization theory.

A particularly simple example is the triangular site percolation problem with $p_c = 1/2$, if we do not divide the lattice into large $b \times b$ blocks but into small triangles of three sites which are nearest neighbors, as shown on the left:

```
   *              x              x
 *   *          x   x          x   .
```

The triangle contains a spanning cluster if either all three sites are occupied (x, central diagram) or two sites are occupied (x) and one site is empty (., right diagram). The first choice appears with probability p^3, the second with probability $p^2(1-p)$. However, this second choice has three possible orientations since each of the three sites can be the single empty site. Thus the total probability of the triangle to have a spanning cluster is

$$p' = p^3 + 3(1-p)p^2 \qquad (13a)$$

with three fixed points p^* where $p' = p$:

$$p^* = 0, \quad p^* = 1/2, \quad p^* = 1\,. \qquad (13b)$$

The second of these fixed points is the percolation threshold, while the first corresponds to lattice animals (Sect. "Small Clusters" and end of Sect. "Fractal Dimension D") and the third to compact non-fractal clusters. With somewhat more effort one can derive also a good approximation for ν.

This agreement of the fixed point p^* with the true threshold $p_c = 1/2$ is not valid for other lattices or block choices. Nevertheless there was a widespread fixed-point consensus that $R_b(p_c) = p_c$ for sufficiently large b. Regrettably, the Evil Empire [20] again destroyed this beauty and found $R_b(p_c) = 1/2$ for square site percolation where $p_c \simeq 0.593$. In general, $R(p_c)$ is a pseudo-universal quantity depending on boundary conditions and sample shape, while p_c for large samples is independent of these details but is different for site and bond percolation and depends on the size of the neighborhood. Life was much nicer before. Fortunately, if a fixed point is determined by Eq. (12c) and the block size goes to infinity, then the fixed point still approaches p_c.

Future Directions

This review summarized the basic theory, particularly when it was not yet contained in the earlier books [1]. Applications were left to the Sahimi book [1]; even for the very first application [2] there is not yet a complete consensus that the three-dimensional percolation exponents apply to polymer gelation. More recent applications are social percolation [21] for marketing by word-of-mouth, and stock market fluctuations due to herding among traders [22].

Percolation theory, similar to Fortran programming or capitalism, was thought to be finished but seems to be alive and kicking. Nevertheless I think the future is more in its applications.

The manuscript was improved by criticism of A. Aharony.

Bibliography

1. Stauffer D (1979) Phys Rep 54:1; Essam JW (1980) Rep Prog Phys 43:843; Stauffer D, Aharony A (1994) Introduction to Percolation Theory. Taylor and Francis, London (revised second edition); Sahimi M (1994) Applications of Percolation Theory. Taylor and Francis, London; Bunde A, Havlin S (1996) Fractals and Disordered Systems. Springer, Berlin; Grimmett G (1999) Percolation, second edition. Springer, Berlin
2. Flory PJ (1941) J Am Chem Soc 63:3083
3. Grassberger P (2003) Phys Rev E 67:036101
4. Redner S (1982) J Statist Phys 29:309
5. Stauffer D, Jan N (2000) In: Khajehpour MRH, Kolahchi MR, Sahimi M (eds) Annual Reviews of Computational Physics, vol VIII (Zanjan School). World Scientific, Singapore
6. Newman MEJ, Ziff RM (2000) Phys Rev Lett 85:4104
7. Tiggemann D (2006) Int J Mod Phys C 17:1141 and Ph D thesis, Cologne University
8. Kasteleyn PW, Fortuin CM (1969) J Phys Soc Jpn Suppl S 26:11
9. Deng VJ, Garoni TM, Sokal AD (2007) Phys Rev Lett 98:030602
10. Nienhuis B (1982) J Phys A 15:199
11. Smirnov S, Werner W (2001) Math Res Lett 8:729
12. Ziff RM (1994) Phys Rev Lett 72:1942
13. Newman CM, Schulman LS (1981) J Stat Phys 26:613
14. de Arcangelis L (1987) J Phys A 20:3057
15. Aizenman M (1997) Nucl Phys (FS) B 485:551
16. Shchur LN, Rostunov T (2002) JETP Lett 76:475
17. Stauffer D (1999) J Irreproducible Results 44:57
18. Jarai AA (2003) Ann Prob 31:444
19. Berger N, Ganten N, Peres Y (2003) Probab Theory Relat Fields 126:221
20. Ziff RM (1992) Phys Rev Lett 69:2670
21. Weisbuch G, Solomon S (2002) In: Bornholdt S, Schuster HG (eds) Handbook of graphs and networks. Wiley-VCH, Weinheim, p 113
22. Cont R, Bouchaud J-P (2000) Macroeconom Dyn 4:170

Scenario-Driven Planning with System Dynamics

NICHOLAS C. GEORGANTZAS
Fordham University Business Schools, New York, USA

Article Outline

Glossary

Mental model how one perceives cause and effect relations in a system, along with its boundary, i. e., exogenous variables, and the time horizon needed to articulate, formulate or frame a decision situation; one's implicit causal map of a system, sometimes linked to the reference performance scenarios it might produce.

Product either a physical good or an intangible service a firm delivers to its clients or customers.

Real option right and obligation to make a business decision, typically a tangible investment. The option to invest, for example, in a firm's store expansion. In contrast to financial 'call' and 'put' options, a *strategic real option* is not tradable. Any time it invests, a firm might be at once acquiring the *strategic real options* of expanding, downsizing or abandoning projects in future. Examples include research and development (abbreviated R&D), merger and acquisition (abbreviated M&A), licensing abroad and media options.

Scenario a postulated sequence or development of events trough time; via Latin *scena* '*scene*', from Greek *σκηνή*, *skēnē* '*tent, stage*'. In contrast to a forecast of what *will* happen in the future, a scenario shows what *might* happen. The term *scenario* must *not* be used loosely to mean situation. *Macro-environmental* as well as *industry-*, *task-* or *transactional-environmental* scenarios are merely inputs to the *strategic objectives* and *real options* a firm must subsequently explore through *strategic scenarios*, *computed* or simulated with an explicit, formal *system dynamics* (abbreviated SD) model of its strategic situation. *Computed strategic scenarios* create the multiple perspectives that strategic thinkers need to defeat the tyranny of dogmatism that often assails firms, governments and other social entities or organizations.

Scenario-driven planning (abbreviated SdP) to attain high performance through strategic flexibility, firms use the SdP *management technology* to create foresight and to anticipate the future with strategic real options, in situations where the business environment accelerates frequently and is highly complex or interdependent, thereby causing uncertainty.

Situation the set of circumstances in which a firm finds itself; its (strategic) state of affairs.

Strategic management process (abbreviated SMP) geared at detecting environmental threats and turning them into opportunities, it *proceeds from* a firm's mission, vision and environmental constraints *to* strategic goals and objectives *to* strategy design or formulation *to* strategy implementation or strategic action *to* evaluation and control *to* learning through feedback (background, Fig. 2).

SMP-1 environmental scanning monitors, evaluates and disseminates knowledge about a firm's internal and external environments to its people. The internal environment contains *s*trengths and *w*eaknesses within the firm; the external shows future *o*pportunities and *t*hreats (abbreviated SWOT).

SMP-2 mission a firm's purpose, *raison d'être* or reason for being.

SMP-3 objectives performance (P) goals that SMP often quantifies for some P metrics.

SMP-4 policy decision-making guidelines that link strategy design or formulation to action or implementation tactics.

SMP-5 strategy a comprehensive plan that shows how a firm might achieve its mission and objectives. The three strategy levels are: corporate, business and process or functional.

SMP-6 strategy design or formulation the interactive, as opposed to antagonistic, interplay of strategic content and process that creates flexible long-range plans to turn future environmental threats into opportunities; includes internal strengths and weaknesses as well as strategic mission and objectives, and policy guidelines.

SMP-7 strategic action or implementation the process by which strategies and policies are put into action through the development of programs, processes, budgets and procedures.

SMP-8 evaluation and control sub-process that monitors activities and performance, comparing actual results with desired performance.

SMP-9 learning through feedback occurs as knowledge about each SMP element enables improving previous SMP elements (background, Fig. 2).

System an organized group of interrelated components, elements or parts working together for a purpose; parts might be either goal seeking or purposeful.

System dynamics (abbreviated SD) a lucid modeling method born from the need to manage business performance through time. Thanks to Forrester [23], who discovered that all change propagates itself through stock and flow sequences, and user-friendly SD software (*iThink*®, *Vensim*®, etc.), SD models let managers see exactly how and why, like other biological and social organizations, business firms perform the way they do. Unlike other social sciences, SD shows exactly how *feedback loops*, i. e., circular cause and effect chains, each containing at least one time lag or delay, interact within a system to determine its performance through time.

Variable or metric something that changes either though time or among different entities at the same time. An *internal change lever* is a decision or policy variable that a strategy-design modeling, or client, team controls. An *external change trigger* is an environmental or policy variable that a strategy-design modeling team does not control. Both *trigger* and *lever* variables can initiate change and be either endogenous or exogenous to a model of a system.

However certain our expectation, the moment foreseen may be unexpected when it arrives
—T.S. Eliot

Definition of the Subject

Many of us live and work in and about business ecosystems with complex structures and behaviors. Some realize that poor performance often results from our very own past actions or decisions, which come back to haunt us. So business leaders in diverse industries and firms, such as *Airbus, General Motors, Hewlett-Packard, Intel* and *Merck*, use scenario-driven planning (SdP) with system dynamics (SD) to help them identify, design and apply high-leverage, sustainable solutions to dynamically complex strategic-decision situations. One must know, for example, if the effect of an environmental change or strategic action gets magnified through time or is dampened and smoothed away. What may seem insignificant at first might cause major disruption in performance. SdP with SD shows the causal processes behind such dynamics, so firms can respond to mitigate impacts on performance.

Accelerating change and complexity in the global business environment make firms and other social organizations abandon their *inactive, reactive* and *preactive* modes [2]. SdP with SD turns them *proactive*, so they can translate anticipation into action. To properly transform anticipation into action, computed with SD models, 'strategic scenarios' must meet four conditions: *consistency, likelihood, relevance* and *transparency* [37]. Combining SdP with SD for that purpose, with other tools, like actor and stakeholder purposes, morphological methods or probability might help avoid entertainment and explore all possible scenarios. Indeed, SdP with SD

> "does not stand alone… modeling projects are part of a larger effort… modeling works best as a complement to other tools, not as a substitute" (see p. 80 in [75]).

SdP with SD is a systematic approach to a vital top-management job: leading today's firm in the rapidly changing and highly complex global environment. Anticipating a world where product life cycles, technology and the mix of collective- and competitive-strategy patterns change at an unprecedented rate is hard enough. Moving ahead of it might prove larger than the talent and resources now available in leading firms. SdP with SD leads to a decisive integration of strategy design and operations, with the dividing line much lower than at present. As mid-level managers take on more responsibility, senior executives become free to give more time and attention to economic conditions, product innovation and the changes needed to enhance creativity toward strategic flexibility [23].

It is perhaps its capacity to reintegrate strategy content and process that turns SdP with SD into a new paradigm for competitive advantage [42], and simulation modeling in general [28], into a critical fifth tool, in addition to the four tools used in science: observation, logical-mathematical analysis, hypothesis testing and experiment [77]. But full-fledged SD models also allow computing scenarios to assess possible implications of strategic situations. Strategic scenarios are not merely hypothesized plausible futures, but computed by simulating combined changes in strategy and in the business environment [32].

Computed scenarios help managers understand what they do not know, enabling strategy design and implementation through the coalignment of timely tactics to improve long-term performance. Through its judicious use of resources, scenario-driven planning with system dynamics makes the tactics required for implementation clear [27]. And because computed scenarios reveal the required coalignment of tactics through time, SdP with SD helps firms become flexible, dependable and efficient, and save time!

Everyone's mind sees differently, but if there is truth in the adage 'a picture is worth a thousand words', then the complex interrelations that SdP with SD unearth and show must be worth billions. In a world where strategic chitchat dominates, one can only hope that SdP with SD will play a central role in public and private dialogues about dynamically complex opportunities and threats.

We shape our buildings; thereafter,
our buildings shape us
—Winston Churchill

Introduction

Following on the heels of Ackoff and Emery [3] and Christensen [10], respectively, Gharajedaghi [35] and Raynor [64] show how strategies with the best chances for brilliant success expose firms to debilitating uncertainty. Firms fail as their recipes for success turn bad through time. Gharajedaghi [35] shows, for example, five strategy scenarios that convert success to failure. Each scenario plays a critically different role. Together, however, these scenarios form a dynamically complex system. Through time, as each scenario plays, it enables the context for the next:

1. *Noble ape* or *copycat* strategy imitates and replicates advantage. Also called 'shadow marketing', it lets shadowy copycats instantly *shadow market* product technology, often disruptively.
2. *Patchy* or *sluggish* strategy delays responses to new technology. When this *second* scenario plays, then patching up wastes time, enabling competitors to deliver new technology and to dominate markets. Worse, it causes costs to rise as it drives down product quality.
3. *Satisficing* or *suboptimal* strategy scenarios take many forms. One entails a false assumption: if a policy lever helps produce desired performance, then pulling or pushing on that lever will push performance further.
4. *Gambling* or *changing the game* strategy scenario transforms a strategic situation by playing the game successfully. While dealing with a challenge, firms gradually transform their strategic situation and change the basis for competition, so a whole new game and set of issues emerge. Success marked, for example, the beginning of the *information era*. But competitive advantage has already moved away from having access to information. In our *systems era* [2], creating new knowledge and generating insight is the new game [81].

Lastly, the cumulative effects from these four strategy scenarios trickle down to the:

5. *Archetypal swing* or *paradigm shift* scenario. Both learning and unlearning can cause archetypal swings and paradigm shifts to unfold through time intentionally [76]. These also occur unintentionally when conventional wisdom fails to explain patterns of events that challenge prevailing mental models. The lack of a convincing explanation creates a twilight zone where acceptable ideas are not competent and competent ideas are not acceptable.

Beliefs about the future drive strategies. But the future is unpredictable. Worse, success demands commitments that make it impossible to adapt to a future that turns out surprising. So, strategies with great success potential also bear high failure probabilities. Raynor [64] calls this the *strategy paradox*. Dissolving it requires turning environmental uncertainty into strategic flexibility. To make it so, Raynor urges managers to: anticipate multiple futures with scenarios, formulate optimal strategies for each future, accumulate *strategic real options* [5] and manage the select options portfolio.

SdP with SD helps managers who operate in an uncertain world question their assumptions about how the world works, so they can see it more clearly. To survive, the human mind overestimates small risks and underestimates large risks. Likewise, it is much more sensitive to losses than to gains. So the capability to leverage opportunities and to mitigate risk might have become an economic value driver.

The purpose of computing scenarios is to help managers alter their view of reality, to match it up more closely with reality as is and as it might become. To become a leader, a manager must define reality. The SdP with SD purpose is *not*, however, to paint a more accurate picture of tomorrow, but to improve the quality of decisions about the future. Raynor says that the requisite strategic flexibility, which SdP with SD creates:

> "is not a pastiche of existing approaches. Integrating these tools and grounding them in a validated theory of organizational hierarchy creates something that is quite different from any of these tools on its own, or in mere combination with the others" (see p. 13 in [64]).

Indeed, knowledge of common purposes and the acceptable means of achieving them form and hold together a purposeful hierarchical system. Its members know and share values embedded in their culture, which knits parts into a cohesive whole. And because each part has a lot to say about the whole, consensus is essential to SdP with SD for the co-alignment of diverse interests and purposes.

Ackoff and Emery [3], Gharajedaghi [35] and Nicolis [55] concur that purpose offers the lens one needs to see a firm as a multi-minded social net. A purposeful firm produces either the same result differently in the same environment or different results in the same or different environments. Choosing among strategic real options is necessary but insufficient for purposefulness. Firms that behave differently but show only one result per environment are goal seeking, not purposeful. Servomechanisms are goal seeking but people are purposeful. As a purposeful system, the firm is part of purposeful sub-systems, such as its *industry value chain* [61] and the society. And firms have purposeful people as members. The result is a dynamically interdependent, i. e., complex, hierarchical purposeful system.

A firm's value chain is, along with its primary and support activities, at once a member of at least one industry value chain and of the society or *macro-environment*. Industry analysis requires looking at value chains independently from the society [61]. But people, the society and firm and industry value chains are so interdependent, so interconnected, that an optimal solution might not exist for any of them independently of the others. SdP with SD helps firms co-align the 'plural rationality' of purposeful stakeholder groups with each other and that of the system as a whole.

Seeing strategic management as a *strategies and tactics net* [27] is in perfect syzygy with the *plural rationality* that SdP with SD accounts for among individuals, groups and organizations. Singer [73,74] contrasts monothematic conventional universes of traditional rationality with the multiverse-directed view of plural rationality. In counterpoint, Morecroft's [52] computed scenarios trace the dysfunctional interactions among sales objectives, overtime and sales force motivation to the intended, i. e., stated, singular rationality that drove action in a large sales organization.

Because their superordinate purpose is neither to compete nor to collaborate, but to develop new wealth-creating capabilities, in unique ways that serve both current and future stakeholder interests, customers and clients included [51], firms can benefit from the multiverse-directed view of strategic management as a net of strategies and tactics. SdP with SD helps firms break free from the tradeoffs tyranny of the mass-production era. Evidently, adherents to tradeoffs-free strategy like *Bell Atlantic*, *Daimler-Benz*, *Hallmark* and *Motorola* "can have it all" [60].

A firm must serve the purposes of its people as well as those of its environment, not as a mindless mechanical system, but as a living, purposeful, *knowledge-bonded* hierarchical system [3,35,55,81]. To clarify, a bike always yields to its rider, for example, regardless of the rider's desire; even if that entails running into a solid brick wall. Ouch! But riding a horse is an entirely different story. Horse and rider form a knowledge-bonded system: the horse must know the rider and the rider must know exactly how to lead the horse.

SdP with SD History: Always Back, Always in Style, Always Practical

Herman Kahn introduced scenarios to planning while at RAND Corporation in the 1950s [45]. Scenarios entered military strategy studies conducted for the US government. In the 1960s, Ozbekhan [58] used urban planning scenarios in Paris, France. Organization theorists and even novelists were quick to catch on. The meaning of scenarios became literary. Imaginative improvisation produced flickering apocalyptic predictions of strikingly optimistic and pessimistic futures. Political and marketing experts use scenarios today to jazz up visions of favorable and unfavorable futures.

Wack [78,79] asserts it was Royal Dutch Shell that came up with the idea of scenarios in the early 1970s. Godet [36] points to the French OTAM team as the first to use scenarios in a futures study by DATAR in 1971. Brauers and Weber [8] claim that Battelle's scenarios method [49] was originally a German approach. In connection with planning, however, most authors see scenario methods as typically American.

Indeed, during the 1970s, US researchers Olaf Helmer and Norman Dalkey developed scenario methods at RAND for eliciting and aggregating group judgments via Delphi and cross-impact matrices [4]. They extended cross impact analysis within statistical decision theory [39]. A synthesis of scenario methods began in the 1970s that draws together multiple views, including those of professional planners, analysts and line managers.

Ansoff [6] and other strategy theorists state that the 1970s witnessed the transformation of global markets. Today, changes in the external sociopolitical environment become pivotal in strategy making. Combined with the geographical expansion of markets, they increase the complexity of managerial work. As environmental challenges move progressively faster, they increase the likelihood of strategic surprises. So, strategic thinkers use scenarios to capture the nonlinearity of turbulent environments. Examples are Hax and Majluf [38] and, more clearly so, Porter [61] and Raynor [64]. They consider scenarios instrumental both in defining uncertainty and in anticipating environmental trends.

Huss and Honton [41] see scenarios emerge as a distinct field of study, a hybrid of a few disciplines. They identify multiple scenarios methods that fall into three major categories:

1. Intuitive logics [78,79], now practiced by *SRI International*,
2. Trend-impact analysis, practiced by the *Futures Group* and
3. Cross-impact analysis, practiced by the *Center for Futures Research* using INTERAX (Interactive Cross-Impact Simulation) and by Battelle using BASICS (BAttelle Scenario Inputs to Corporate Strategies).

Similarly, after joining Ozbekhan to advocate reference scenarios, Ackoff [2] distinguishes between:

1. Reference projections as piecemeal extrapolations of past trends and
2. The overall reference scenario that results from putting them together.

Based on Acar's [1] work under Ackoff, Georgantzas and Acar [32] explore these distinctions with a practical managerial technology: *comprehensive situation mapping* (CSM). CSM is simple enough for MBA students to master in their capstone Business Policy course. With the help of *Vensim® PLE* [18], CSM computes scenarios toward achieving a well-structured process of managing ill-structured strategic situations. In their introduction to SD, Georgantzas and Acar (see Chap. 10 in [32]) draw from the banquet talk that Jay Wright Forrester, Germeshausen Professor Emeritus, MIT, gave at the 1989 *International Conference of the System Dynamics Society*, in Germany, at the University of Stuttgart:

After attending the Engineering College, University of Nebraska, which included control dynamics at its core, Forrester went to MIT. There he worked for Gordon S. Brown, a pioneer in feedback control systems. During World War II, Brown and Forrester worked on servomechanisms for the control of radar antennas and gun mounts. This was research toward an extremely practical end, during which Forrester run literally from mathematical theory to the battlefield, aboard the US carrier *Lexington*.

After the war, Forrester worked on an analog aircraft flight simulator that could do little more than solve its own internal idiosyncrasies. So, Forrester invented *random-access magnetic storage* or *core memory*. His invention went into the heart of *Whirlwind*, a digital computer used for experimental development of military combat systems that eventually became the *s*emi*a*utomatic *g*round *e*nvironment (SAGE) air defense system for North America.

Alfred P. Sloan, the man who built *General Motors*, founded the *Sloan School of Management* in 1952. Forrester joined the school in 1956. Having spent fifteen years in the science and engineering side of MIT, he took the challenge of exploring what engineering could do for management.

One day, he found himself among students from *General Electric*. Their household appliance plants in Kentucky puzzled them: they would work with three or four shifts for some time and then, a few years later, with half the people laid off. Even if business cycles would explain fluctuating demand, that did not seem to be the entire reason. *GE*'s managers felt something was wrong.

After talking with them about hiring, firing and inventory policies, Forrester did some simulation on a paper pad. He started with columns for inventories, employees and customer orders. Given these metrics and *GE*'s policies, he could tell how many people would be hired or fired a week later. Each decision gave new conditions for employment, inventories and production. It became clear that wholly determined internally, the system had potential for oscillatory dynamics. Even with constant incoming orders, the policies caused employment instability. That longform simulation of *GE*'s inventory and workforce system marked the beginning of system dynamics [23,24,25,26].

SdP with SD Use and Roadmap

Scenarios mostly help forecast alternative futures but, as firms abandon traditional forecasting methods for interactive planning systems, line managers in diverse business areas adopt scenario-driven planning with system dynamics. Realizing that a tradeoffs-free strategy design requires insight about a firm's environment, both business and sociopolitical, to provide intelligence at *all* strategy levels, firms use SdP with SD to design *corporate, business* and *process* or *functional* strategies. SdP with SD is not a panacea and requires discipline, but has been successful in many settings. Its transdisciplinary nature helps multiple applications, namely capital budgeting, career planning, civil litigation [31], competitive analysis, crisis management, decision support systems (DSS), macroeconomic analysis, marketing, portfolio management and product development [65]. SdP with SD is a quest for managers who wish to be leaders, not just conciliators. They recognize that *logical incrementalism*, a piecemeal approach, is inadequate when the environment and their strategy change together.

Top management might see both divisional, i. e., business, and process or functional strategies as ways of implementing corporate strategy. But *active subsidiaries* [43,44] provide both strategic ideas and results to their parent enterprise. Drawing too stiff a line between the corporate office and its divisions might be

> "an unhealthy side effect of our collective obsession with generating returns. The frameworks for developing competitive strategy that have emerged over the last thirty years have given us unparalleled insight into how companies can succeed. And competitive strategy remains enormously important, but it should be the preserve of divisional management… corporate strategy should be focused on the management of strategic uncertainty" (see p. 11 in [64]).

Roadmap It is material to disconnect scenarios from unproductive guesswork and to anchor them to sound practices for strategy design. This guided tour through the fascinating but possibly intimidating jungle of scenario definitions shows what the future might hold for SdP with SD. Extensive literature, examples, practical guidelines and two real-life cases show how computed scenarios help manage uncertainty, that necessary disciple of our open market system. Unlike extrapolation techniques, SdP with SD encourages managers to think broadly about the future.

The above sections clarify the required context and provide a glossary. Conceptual confusion leads to language games at best and to operational confusion at worst [15]. SdP with SD helps firms avert both types of confusion. Instead of shifting their focus away from actuality and rationality, managers improve their insight about fundamental assumptions underlying changes in strategy. The mind-set of SdP with SD makes it specific enough to give practical guidance to those managing in the real world, both now and in the future.

The sections below look at *three* SdP with SD facets linked to strategy design and implementation. The *first* facet involves the business environment, the forces behind its texture and future's requisite uncertainty (Sect. "Environmental Turbulence and Future Uncertainty"). The *second* entails unearthing unstated assumptions about changes in the environment and in strategy, and about their potential combined effects on performance. The SdP with SD framework (Sect. "SdP with SD: The Modeling Process ≡ Strategic Situation Formulation") builds on existing scenario methods. Its integrative view delineates processes that enhance institutional learning, bolster productivity and improve performance through strategic flexibility. It shows why interest in computed scenarios is growing.

The *third* facet entails *computing* the combined or mixed effects on performance of changes both in the environment and in strategy. Even in mature economies, no matter how and how frequently said, decision makers often forget how the same action yields different results as the environment changes. The result is often disastrous. Conversely, the tight coupling between computed scenarios and strategic results can create new knowledge. Linking a mixed environmental and decision scenario in a one-to-one correspondence to a strategic result suits the normative inclination of strategic management, placing rationalistic inquiry at par with purely descriptive approaches in strategy research.

The unified treatment of SdP with SD and the strategy-making process grants a practical bonus, accounting for the entry's peculiar nature. It is not only a conceptual or idea contribution, but also an application-oriented entry. Sections "Case 1: Cyprus' Environment and Hotel Profitability" and "Case 2: A Japanese Chemicals Keiretsu (JCK) present two real-life cases of scenario-driven planning with system dynamics. Written with both the concrete and the abstract thinker in mind, the two cases show how firms and organizations build scenarios with a modest investment. SdP with SD provides an effective management technology that serves well those who adopt it. It saves them both time and energy.

Improvements in causal mapping [19,20], and SD modeling and analysis [50,57] contribute to the SdP with SD trend (Sect. "Future Directions"). Behavioral decision theory and cognitive science also help translate the knowledge of managers into SD models. The emphasis remains on small, transparent models of strategic situations and on dialogue between the managers' mental models and the computed scenarios [53].

All prognosticators are bloody fools
—Winston Churchill

Environmental Turbulence and Future Uncertainty

Environmental Turbulence

Abundant frameworks describe the business environment, but the one by Emery and Trist [22], which Duncan [17] abridged, has been guiding many a strategic thinker. It shows four business environments, each more complex and troublesome for the firm than the preceding one (Fig. 1a).

1. *Placid* or *independent-static environment*: infrequent changes are independent and randomly distributed,

i. e., IID. Surprises are rare, but no new major opportunities to exploit either (*cell* 1, Fig. 1a).

2. *Placid-clustered* or *complex-static environment*: patterned changes make forecasting crucial. Comparable to the economist's idea of imperfect competition, this environment lets firms develop distinctive competencies to fit limited opportunities that lead to growth and bureaucracy (*cell* 2, Fig. 1a).
3. *Disturbed-reactive* or *independent-dynamic environment*: firms might influence patterned changes. Comparable to oligopoly in economics, this environment makes changes difficult to predict, so firms increase their operational flexibility through decentralization (*cell* 3, Fig. 1a).
4. *Turbulent field* or *complex-dynamic environment*: most frequent, changes are also complex, i. e., interdependent, originating both from autonomous shifts in the environment and from interdependence among firms and conglomerates. Social values accepted by members guide strategic response (*cell* 4, Fig. 1a).

Ansoff and McDonnell [7] extend the dichotomous environmental uncertainty perceptions by breaking turbulent environments (*cell* 4, Fig. 1a) into *discontinuous* and *surprising*. This is a step in the right direction, but not as helpful as a causal model specific to the system structure of a firm's strategic situation. Assuredly, 2 × 2 typologies help clarify exposition and are most frequent in the organization theory and strategy literatures. The mystical significance of duality affected even Leibniz, who associated one with God and zero with nothingness in the binary system. The generic solutions that dichotomies provide leave out the specifics that decision makers need. No matter what business they are in (Fig. 1b), managers cannot wait until a better theory comes along; they must act now.

It is worth noting that people often confuse the term 'complex' with 'complicated'. Etymology shows that *complicated* uses the Latin ending *-plic*: *to fold*, but *complex* contains the Greek root *πλέξ-* '*plēx-*': *to weave*. A complicated structure is thereby folded, with hidden facets stuffed into a small space (Fig. 1c). But a complex structure has interwoven parts with mutual interdependencies that cause dynamic complexity [46]. Remember: complex is the opposite of independent or untwined (Fig. 1a) and complicated is the opposite of simple (Fig. 1c).

Daft and Weick's [12] vista on firm *intrusiveness* and *environmental equivocality* is pertinent here. They see many events and trends in the environment as being inherently unclear. Managers discuss such events and trends, and form mental models and visions expressed in a fuzzy language and label system [80]. Within an *enactment* process, equivocality relates to managerial assumptions underlying the *analyzability* of the environment. A firm's *intrusiveness* determines how *active* or *passive* the firm is about environmental scanning. In this context, as the global environment gets turbulent, active firms and their subsidiaries construct SdP with SD models and compute scenarios to improve performance.

Managers of active firms combine knowledge acquisition with interpretations about the environment and their strategic situation. They reduce equivocality by assessing alternative futures through computed scenarios. In frequent meetings and debates, some by videoconferencing, managers use the dialectical inquiry process for *s*trategic *a*ssumption *s*urfacing and *t*esting (SAST), a vital strategic loop. Often ignored, the SAST loop gives active firms a strategic compass [47].

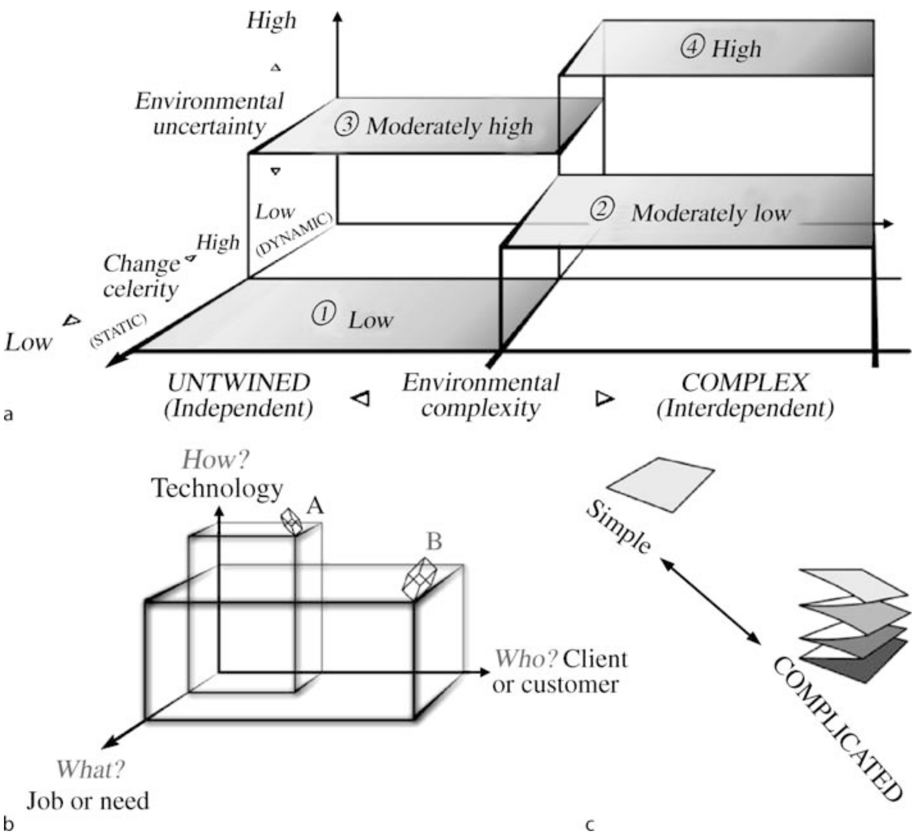
Conversely, passive firms do not actively seek knowledge but reduce equivocality through rules, procedures and regular reports: reams of laser-printed paper with little or no pertinent information. Managers in passive firms use the media to interpret environmental events and trends. They obtain insight from personal contacts with significant others in their environment. Data are personal and informal, obtained as the opportunity arises.

Future Uncertainty

"If we were omnipotent", says Ackoff, then we could get "perfectly accurate forecasts" (see p. 60 in [2]). Thank God the future is unpredictable and we must yet create it. If it were not, then life would have been so boring! Here are some facts about straight forecasting:

1. Forecasts are seldom perfect, in fact, they are always wrong, so a useful forecasting model is one that minimizes error.
2. Forecasts always assume underlying stability in systems.
3. Product family and aggregated forecasts are always more accurate than single product forecasts, so the large numbers law applies.
4. In the short-term, managers can forecast but cannot act because time is too short; in the long term, they can act but cannot forecast.

To offset conundrum #4, SdP with SD juxtaposes the decomposition of performance dynamics into the growth and decline archetypes caused by *balancing* (–) and *reinforcing* (+) recursive causal-link chains or *feedback loops* [33,50]. A thermostat is a typical example of a goal-seeking feedback loop that causes either balancing growth



Scenario-Driven Planning with System Dynamics, Figure 1
a Environmental complexity and change celerity dimensions that cause perceived environmental uncertainty (adapted from [32]). b Scenario-driven planning with system dynamics helps with strategy-design fundamentals, such as, for example, defining a business along the requisite client-job-technology three-dimensional grid. c The simple-complicated dimension must *not* be confused with the environmental complexity dimension (adapted from [46])

or decline. The gap between desired and room temperature causes action, which alters temperature with a time lag or delay. Temperature changes in turn close the gap between desired and room temperature.

Conversely, a typical loop that feeds on itself to cause either exponential growth or decline is that of an arms race. One side increases its arms. The other sides increase theirs. The first side then reacts by increasing its arms, and so on. Price wars between stores, promotional competition, shouting matches, one-upmanship and the wildcard interest rates of the late 1970s are good examples too. Escalation might persist until the system explodes or outside intervention occurs or one side quits, surrenders or goes out of business. In the case of wildcard interest rates, outside intervention by a regulatory agency can bring an end to irrationally escalating rates.

We've never been here before
—Peter Senge

SdP with SD: The Modeling Process ≡ Strategic Situation Formulation

The strategic management process (SMP, Fig. 2) starts with environmental scanning, in order to gauge environmental trends, opportunities and threats. Examples include increasing rivalry among existing competitors and Porter's [62] emphasis on the bargaining power of buyers and suppliers as well as on the threats of new entrants and substitutes. Even if some firms reduce environmental scanning to industry analysis in practice, changes in the environment beyond an industry's boundaries can determine what happens within the industry and its entry, exit and inertia barriers. Internal capability analysis comes next. It examines a firm's past actions and internal policy levers that can both propel and limit future actions. The integrative perspective of the SdP with SD framework on Fig. 2 delineates processes that enhance institutional learning, bolster productivity and improve performance through strategic flexibility.

Strategy design begins by identifying variables pertinent to a firm's strategic situation, along with their interrelated causal links. Changes in these variables can have profound effects on performance. Some of the variables belong to a firm's external environment. Examples are emerging new markets, processes and products, government regulations and international interest and currency rates. Changes either in these or their interrelated causal links determine a firm's performance through time.

It is a manager's job to understand the causal links underlying a strategic situation. SdP with SD helps anticipate the effects of future changes triggered in the external environment. Other variables are within a firm's control. Pulling or pushing on these internal levers also affects performance. To evaluate a change in strategy, one must look at potential results along with changes in the environment, matching resource capabilities, stakeholder purposes, and organizational goals and objectives (Fig. 2).

Most variables interact. Often, the entire set of possible outcomes is obscure, difficult to imagine. But if managers oversimplify, then they end up ignoring the combined effects of chain reactions. Even well-intended rationality often leads to oversimplification, which causes cognitive biases (CBs) that mislead decision makers [21,70,72]. Conversely, computing mixed environmental and decision scenarios that link internal and external metrics can reveal unwarranted simplification.

SdP with SD integrates business intelligence with strategy design, not as a narrow specialty, but as an admission of limitations and environmental complexity. It also uses multiperspective dialectics, crucial for strategic assumption surfacing and testing (SAST). Crucial because the language and labels managers use to coordinate strategic real options are imprecise and fuzzy. Fuzzy language is not only adequate *initially* for managing interdependence-induced uncertainty but required [80]. Decision makers rely on it to overcome psychological barriers and Schwenk's [70] groups of CBs.

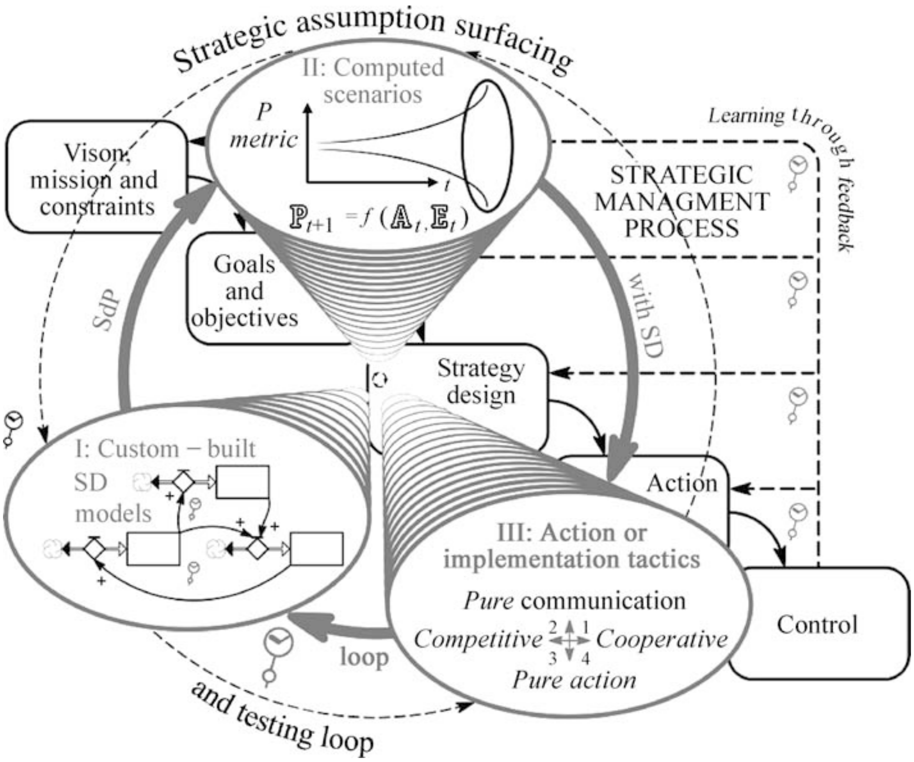
The best-case scenario for a passive firm is to activate modeling on Fig. 2, sometimes unknowingly. When its managers boot up, for example, electronic spreadsheets that contain inside-out causal models, with assumptions hidden deeply within many a formula. At bootup, only the numbers show. So passive-firm managers use electronic spreadsheets to laser-print matrices with comforting numbers. They

> "twiddle a few numbers and diligently sucker themselves into thinking that they're forecasting the future" [69].

And that is only when rapid changes in the environment force them to stop playing *blame the stakeholder*. They stop fighting the last war for a while, artfully name the situation a crisis, roll up their sleeves, and chat about and argue, but quickly agree on some arbitrary interpretation of the situation to generate strategic face-saving options. Miller and Friesen (see pp. 225–227 in [48]) show how for futile firms, rapid environmental changes lead to crisis-oriented decisions. Conversely, successful firms look far into the future as they counter environmental dynamism through strategy design with real options. Together, their options and interpretation of the environment, through the consensus that SdP with SD facilitates, enable a shared logic to emerge: a shared mental model that filters hidden spreadsheet patterns and heroic assumptions clean and clear.

Managers of active firms enter the SdP with SD loop of Fig. 2 both consciously and conscientiously. They activate strategic intelligence via computed scenarios and the SAST loop. Instead of twiddling spreadsheet numbers, *pro*active firm managers twiddle model assumptions. They stake, through SD model diagrams, their intuition about how they perceive the nature and structure of a strategic situation. Computed scenarios quantitatively assess their perceived implications. Having quantified the implications of shared visions and claims about the structure of the strategic situation, managers of active firms are likely to reduce uncertainty and equivocality. Now they can manage strategic interdependence. Because articulated perception is the starting point of all scenarios, computed scenarios give active firms a fair chance at becoming fast strategic learners.

The design of action or implementation tactics requires detailing how, when and where a strategy goes into action. In addition to assuming the form of *pure communication* (III: 1 and 2, Fig. 2) or *pure action* (III: 3 and 4, Fig. 2), in a pragmatic sense, tactics can be either cooperative or competitive and defensive or offensive. Market location tactics, for example, can be either offensive, trying to rob market share from established competitors, or defensive, preventing competitors from stealing one's market share. An offensive tactic takes the form of frontal assault, flanking maneuver, encirclement, bypass attack or guerilla warfare. A defensive tactic might entail raising structural barriers, increasing expected retaliation or lowering the inducement for future attack. Conversely, cooperative tactics try to gain mutual advantage by working with rather than against others. Cooperative tactics take the form of alliances, joint ventures, licensing agreements, mutual service consortia and value-chain partnerships, the co-location of which often creates industrial districts [29].


Strategic assumption surfacing
II: Computed scenarios
P metric
t
$\mathbb{P}_{t+1} = f(\mathbb{A}_t, \mathbb{E}_t)$
Learning through feedback
Vison, mission and constraints
STRATEGIC MANAGMENT PROCESS
Goals and objectives
SdP
with SD
Strategy design
I: Custom – built SD models
+
Action
III: Action or implementation tactics
Pure communication
Competitive 2 1 Cooperative 3 4
Pure action
Control
loop
and testing loop

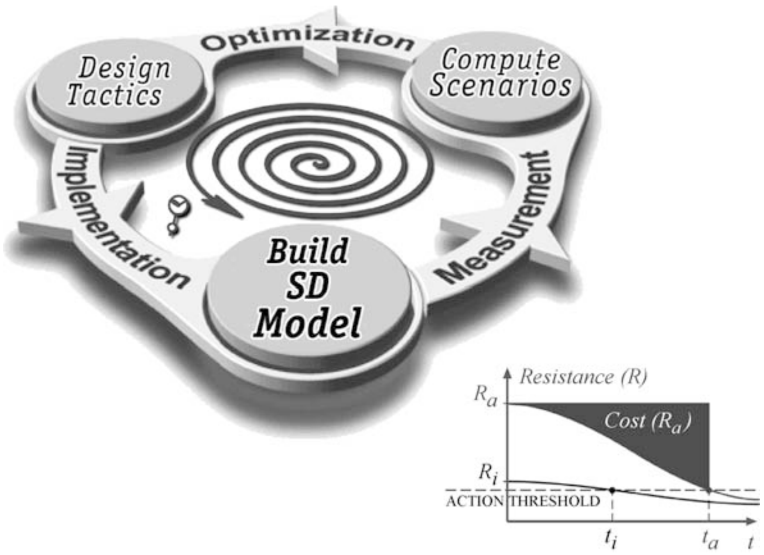
Scenario-Driven Planning with System Dynamics, Figure 2
Cones of resolution show how scenario-driven planning with system dynamics enhances the strategy design component of the strategic management process (SMP; adapted from [32])

The usual copycat strategy retort shows linear thinking at best and clumsy *benchmarking*, also known as shadow marketing, at worst. Its proponents assume performance can improve *incrementally*, with disconnected tactics alone, when strategy design is of primary concern. Piecemeal tactics can undermine strategy, but they are secondary. It might be possible to improve performance through efficient tactics, but is best to design strategies that expel counterproductive tactics. Counterproductive tactics examples are coercive moves that increase rivalry, without a real payoff, either direct or indirect, for the industry incumbent who initiates them. It is atypical of an industry or market leader to initiate such moves.

In strategy, superb action demands superior design. According to the design school, which Ansoff, Channon, McMillan, Porter, Thomas and others lead, logical incrementalism may help implementation, but becomes just another prescription for failure when the environment shifts. Through its judicious use of corporate resources, SdP with SD makes the tactics required for action clear. Also, it reveals their proper coalignment through time, so a firm can build strategic flexibility and save time!

The Modeling Process ≡ Strategic Situation Formulation

SdP with SD (Fig. 2) begins by modeling a business or 'social process' than a business or 'social system'. It is more productive to identify a *social process* first and then seek its causes than to slice a chunk of the real world and ask what dynamics it might generate. Distinguishing between a *social system* and a social process is roughly equivalent to distinguishing between a system's underlying causal structure and its dynamics. Randers (see p. 120 in [63]) defines a social system as a set of cause and effect relations. Its structure is a causal diagram or map of a real-world chunk. A social process is a behavior pattern of events evolving through time. The simulation results of SdP with SD models show such chains of events as they might occur in the real world. An example of a social system (structure) is the set of rules and practices that a firm might enact when dealing with changes in demand, along with the communication channels used for transmitting information and decisions. A corresponding social process (dynamics) might be the stop-and-go pattern of capital in-



Scenario-Driven Planning with System Dynamics, Figure 3
The recursive nature of the modeling process that scenario-driven planning with system dynamics entails a creates a sustainable, ever-expanding vortex of insight and wisdom, needed in strategic real-options valuation, and b saves both time and money as it renders negligible the cost of resistance (*R*) to change

vestment caused by a conservative bias in a firm's culture.

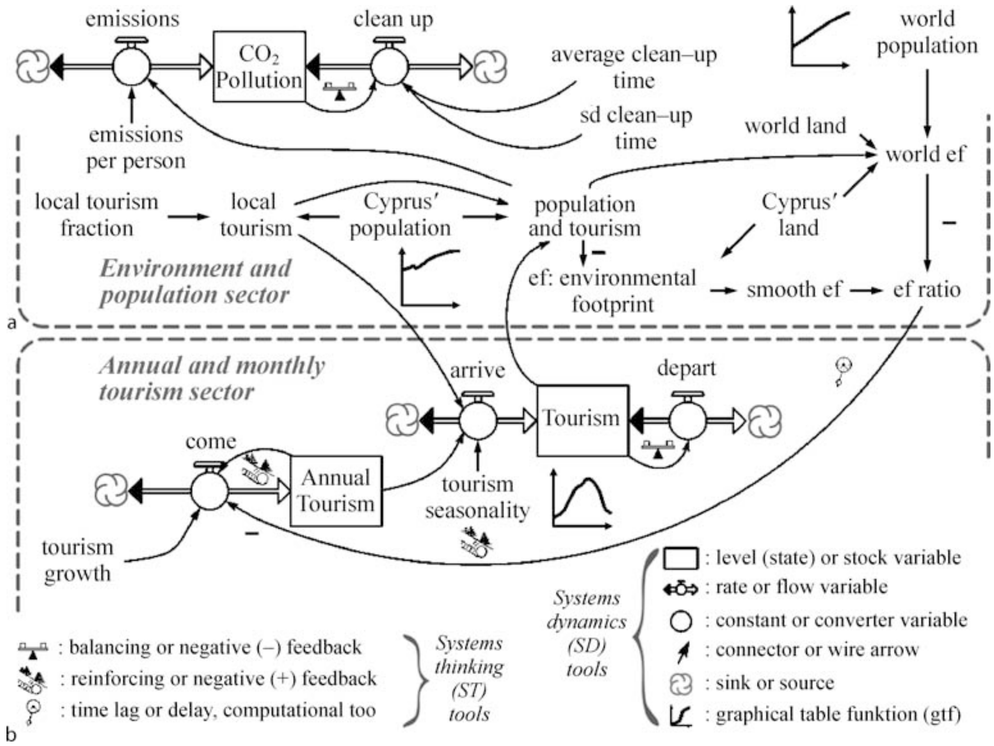
In his model of a new, fast-growing product line, for example, Forrester [24] incorporates such a facet of corporate culture. Causing sales to stagnate, considerable back orders had to accumulate to justify expansion because the firm's president insisted on personally controlling all capital expenditures.

People often jump into describing system structure, perhaps because of its tangible nature as opposed to the elusive character of dynamics or social process fragments. Also, modelers present model structure first and then behavior. Ultimately, the goal in modeling a strategic situation is to link system structure and behavior. Yet, in the early stages of modeling is best to start with system dynamics and then seek underlying causes. Indeed, SD is particularly keen in understanding system performance, "not structure per se" (see p. 331 in [56]), in lieu of SD's core tenet that structure causes performance.

The modeling process itself is recursive in nature. The path from real-world events, trends and negligible externalities to an effective formal model usually resembles an expanding spiral (Fig. 3a). A useful model requires conceptualization; also focusing the modeling effort by establishing both the time horizon and the perspective from which to frame a decision situation. Typically, strategy-design models require a long-term horizon, over which computed scenarios assess the likely effects of changes both in strategy and in the environment.

Computer simulation is what makes SdP with SD models most useful. Qualitative cause and effect diagrams are too vague, tricky to simulate mentally. Produced through knowledge elicitation, their complexity vastly exceeds the human capacity to see their implications. Casting a chosen perspective into a formal SdP with SD model entails postulating a detailed structure; a diagramming description precise enough to propagate images of alternative futures, i. e., computed scenarios, "though not necessarily accurate" (see p. 118 in [63]). But the modeling process must never downplay the managers' mental database and its knowledge content. Useful models always draw on that mental database [24].

Following Morecroft [53], SdP with SD adopters might strive to replace the notion of modeling an objectively singular world *out there*, with the much softer approach of building formal models to improve managers' mental models. The expanding spiral of Fig. 3a shows that the insight required for decisive action increases as the quantity of information decreases, by orders of magnitude. The required quantification of the relations among variables pertinent to a strategic situation changes the character of the information content as one moves from mental to written to numerical data. Perceptibly, a few data remain, but much more pertinent to the nature and structure of the situation. Thanks to computed scenarios, clarity rules in the end. And, if the modeling process stays *i*nteractive (i), as opposed to *a*ntagonistic (a), then clarity means low resistance to change ($R_i < R_a$, Fig. 3b), which helps reach


emissions
clean up
CO_2 Pollution
average clean–up time
sd clean–up time
world population
emissions per person
world land
world ef
local tourism fraction
local tourism
Cyprus' population
population and tourism
Cyprus' land
–
Environment and population sector
ef: environmental footprint
smooth ef
ef ratio
a
Annual and monthly tourism sector
arrive
depart
Tourism
come
Annual Tourism
tourism seasonality
–
tourism growth
: level (state) or stock variable
: rate or flow variable
: constant or converter variable
: connector or wire arrow
: sink or source
: graphical table funktion (gtf)
Systems dynamics (SD) tools
: balancing or negative (–) feedback
: reinforcing or negative (+) feedback
: time lag or delay, computational too
Systems thinking (ST) tools
b

Scenario-Driven Planning with System Dynamics, Figure 4
Cyprus' a environment and population, and b annual and monthly tourism model sectors (adapted from and extending [30])

a firm's action/implementation threshold quickly ($t_i < t_a$, Fig. 3b). This is how SdP with SD users build strategic flexibility while they save both time and money!

Case 1: Cyprus' Environment and Hotel Profitability

Cyprus' Hotel Association wished to test how Cyprus' year 2010 official tourism strategy might affect tourist arrivals, hotel bed capacity and profitability, and the island's environment [30]. Computed with a system dynamics simulation model, four tourism growth scenarios show what might happen to Cyprus' tourism over the next 40 years, along with its potential effects on the sustainability of Cyprus' environment and hotel profitability. Following is a partial description of the system dynamics model that precedes its dynamics.

Model Description (Case 1)

The SD model highlights member interactions along Cyprus' hotel value chain. The model incorporates a generic value-chain management structure that allows modeling customer-supplier value chains in business as well as in physical, biological and other social systems. Although the structure is generic, its situation specific parameters faithfully reproduce the dynamic behavior patterns seen in Cyprus' hotel value-chain processes, business rules and resources.

Cyprus' Environment, Population and Tourism Model Sectors Within Cyprus' environment and population sector (Fig. 4a), the carbon dioxide (CO_2) pollution stock is the accumulation of Cyprus' anthropogenic emissions less the Mediterranean Sea region's self clean-up rate. The clean-up rate that drains Cyprus' CO_2 pollution depends on the level of anthropogenic pollution itself as well as on the average clean-up time and its standard deviation (sd). Emissions that feed CO_2 pollution depend on Cyprus' population and tourism and on emissions per person [9].

In SD models, rectangles represent stocks, i. e., level or state variables that accumulate through time, e. g., the Tourism stock on Fig. 4b. The double-line, pipe-and-valve-like icons that fill and drain the stocks, often emanating from cloud-like sources and ebbing into cloud-like sinks, represent material flows that cause the stocks to change. The arrive rate of Fig. 4b, for example, shows tourists who flow into the tourism stock per month. Single-line arrows represent information flows, while plain text or circular icons depict auxiliary constant or converter variables, i. e., behavioral relations or decision points that convert information into decisions. Changes in the tourism stock, for example, depend on annual tourism, adjusted by tourism seasonality. Both the diagram on Fig. 4a and Table 1 are reproduced from the actual simulation model, first built on the glass of a computer screen using the diagramming interface of *iThink*® [67], and then specifying simple algebraic equations and parameter values. Built-in functions help quantify policy parameters and variables pertinent to Cyprus' tourism situation.

There is a one-to-one correspondence between the model diagram on Fig. 4a and its equations (Table 1). Like the diagram, the equations are the actual output from *iThink*® too. The equations corresponding to Fig. 8b are archived in [30]. Together, Cyprus' population, local tourism and monthly tourism determine the population and tourism sum (Eq. 1.11, Table 1). According to CYSTAT [11], both *Cyprus' Tourism Organization* and its government attach great importance to local tourism. A study on domestic tourism conducted in 1995 revealed that about 46 percent of Cypriots take long holidays. Of these, 61 percent take long holidays exclusively in Cyprus and eight percent in Cyprus and abroad, while 31 percent chose to travel abroad only. These are precisely the percentages in the model (Eq. 1.10, Table 1).

On Fig. 4a, the world land and population data, minus Cyprus' land, population and tourism co-determine the world EF (environmental footprint, Eq. 1.14, Table 1). Compared to Cyprus' smooth EF, i. e., the smooth ratio of the island's free land divided by its total population and tourism, the world EF gives a dynamic measure of Cyprus' relative attractiveness to the rest of the world. The EF ratio (Eq. 1.6, Table 1), i. e., the ratio of Cyprus' smooth EF (Eq. 1.13, Table 1) divided by the world EF (Eq. 1.14, Table 1), assumes that the higher this ratio is, the more attractive the island is to potential tourists, and vice versa. The EF ratio, which depends on Cyprus' total population and tourism, feeds back to the island's annual tourism via the inflow of foreign visitors who come to visit Cyprus every year (Fig. 4b).

The logistic or Verhulst growth model, after François Verhulst who published it in 1838 [66], helps explain Cyprus' actual annual tourism, a quantity that cannot grow forever (Fig. 4b). Every system that initially grows exponentially eventually approaches the carrying capacity of its environment, whether it is food supply for moose, the number of people susceptible to infection or the potential market for a good or a service. As an 'autopoietic' system approaches its limits to growth, it goes through a non-linear transition from a region where positive feedback dominates to a negative feedback dominated regime. S-shaped growth often results: a smooth transition from exponential growth to equilibrium.

The logistic model conforms to the requirements for S-shaped growth and the ecological idea of carrying capacity. The population it models typically grows in a fixed environment, such as Cyprus' foreign annual tourism has done since 1960 up to 2000. Initially dominated by positive feedback, Cyprus' annual tourism might soon reach the island's carrying capacity, with a nonlinear shift to dominance by negative feedback. While accounting for Cyprus' tourism lost to the summer of 1974 Turkish invasion, officially a very long 'military intervention', further depleting annual tourism is the outflow of Cyprus' visitors (not shown here) who might go as the island's free area reaches its Carrying Capacity, estimated at seventy times the number of Cyprus' visitors in 1960 [30].

Cyprus' Hotel Association listed Cyprus tourism seasonality as one of its major concerns. At the time of this investigation, CYSTAT [11] had compiled monthly tourism data for only 30 months. These were used for computing Cyprus' tourism seasonality (Fig. 4b). Incorporating both the foreign annual tourism and the monthly tourism stocks in the model allows both looking at the big picture of annual tourism growth and assessing the potential long-term effects of tourism seasonality on the sustainability of Cyprus' environment and hotel EBITDA, i. e., *e*arnings *b*efore *i*nterest, *t*axes, *d*epreciation and *a*mortization. The publicly available actual annual tourism data allow testing the model's usefulness, i. e., how faithfully it reproduces actual data between 1960 and 2000 [30].

Cyprus' foreign visitors and local tourists arrive at the island's hotels and resorts according to Cyprus' tourism seasonality, thereby feeding Cyprus' monthly tourism stock. About 11.3 days later, according to CYSTAT's [11] estimated average stay days, both foreign visitors and local tourists depart, thereby depleting the monthly tourism stock. By letting tourism growth = 0 and Cyprus' tourism seasonality continue repeating its established pattern, the model computes a zero-growth or *base-run* scenario. Subsequently, however, tourism growth values other than zero initiate different scenarios.

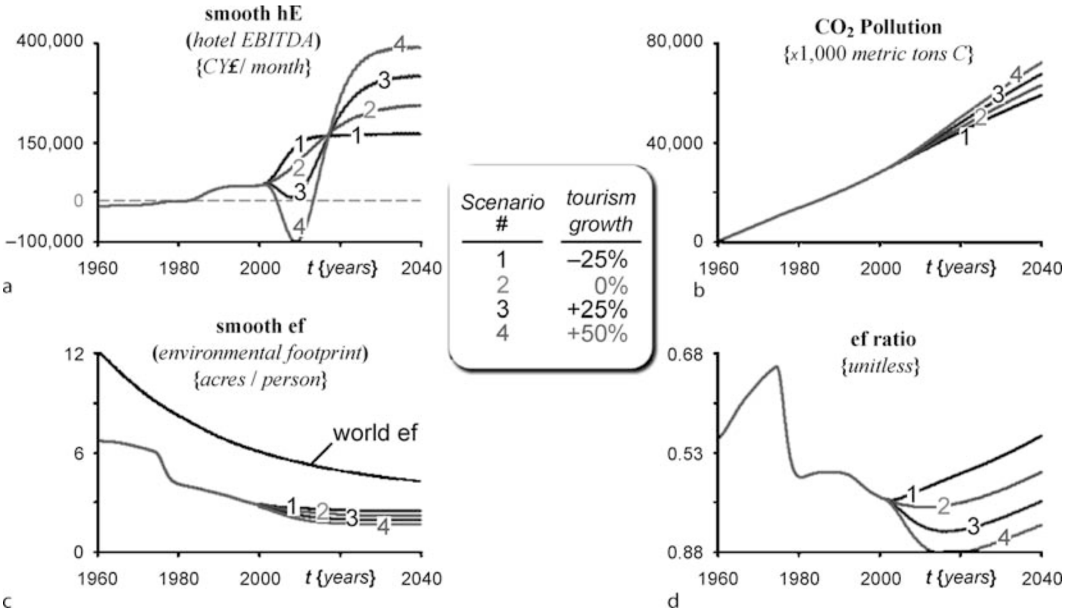
Cyprus' Tourism Growth Scenarios (Case 1)

What can Cyprus' hoteliers expect to see in terms of bottom-line dynamics? According to the four tourism-growth scenarios computed on Fig. 5, seasonal variations notwithstanding, the higher Cyprus' tourism growth is, the lower hotel EBITDA (smooth hE) is, in the short term. In the long term, however, higher tourism growth yields higher profitability in constant year 2000 prices.

Scenario-Driven Planning with System Dynamics, Table 1
Cyprus' environment and population (and local tourism) model sector (Fig. 4a) equations, with variable, constant parameter and unit definitions

Level or state variables **(stocks)**	***Eq.*** #
$CO_2Pollution(t) = CO_2Pollution(t - dt) + (emissions - clean\ up) * dt$	(1.1)
INIT CO_2 Pollution = emissions (Based on 1995 gridded carbon dioxide anthropogenic emission data; unit: 1000 metric ton C per one degree latitude by one degree longitude grid cell)	(1.1.1)
Rate variables **(flows)**	
Emissions = emissions per person ∗ population and tourism (unit: 1000 metric tons C/month)	(1.2)
Cleanup = max(0, CO_2Pollution/average clean − up time) (unit: 1000 metric tons C/month)	(1.3)
Auxiliary variables and constants **(converters)**	
Average clean − up time = 1200 (Med Sea region average self clean-up time = 100 years; unit: months)	(1.4)
Cyprus' land = If (time ≤ 168) then (9251 ∗ 247.1052) else ((9251 − 3355) ∗ 247.1052) (Cyprus' free land area; unit: acres; 1 km^2 = 247.1052 acres)	(1.5)
EF ratio = smooth EF/world EF (unit: unitless)	(1.6)
EF : environmental footprint = Cyprus' land/population and tourism (unit: acres/person)	(1.7)
Emissions per person = 1413.4/702000/12 (unit: anthropogenic emissions/person/month)	(1.8)
Local tourism = local tourism fraction ∗ Cyprus' population (unit: persons/month)	(1.9)
Local tourism fraction = 0.46 ∗ (0.61 + 0.08) (Percentages based on a 1995 study on domestic tourism; unit: unitless)	(1.10)
Population and tourism = Cyprus' population + Tourism − local tourism (Subtracts local tourists already included in Cyprus' population; unit: persons)	(1.11)
Sd clean − up time = 240 (clean-up time standard deviation = 20 years; unit: months)	(1.12)
Smooth EF = SMTH3 (EF: environmental footprint, 36) (Third-order exponential smooth of EF)	(1.13)
World EF = (world land − Cyprus' land)/(world population − population and tourism) (unit: acres/person)	(1.14)
World land = 36677577730.80 (unit: acres)	(1.15)
Cyprus' population = GRAPH(time/12) (Divided by 12 since these are annual data; unit: persons) (0.00, 493984), (1.00, 498898), (2.00, 496570), (3.00, 502001), (4.00, 505622), (5.00, 509329), (6.00, 512950), (7.00, 516743), (8.00, 520968), (9.00, 525364), (10.0, 529847), (11.0, 534330), (12.0, 539934), (13.0, 546486), (14.0, 552348), (15.0, 526313), (16.0, 516054), (17.0, 515881), (18.0, 518123), (19.0, 521657), (20.0, 526744), (21.0, 532692), (22.0, 538210), (23.0, 544675), (24.0, 551659), (25.0, 558038), (26.0, 560366), (27.0, 568469), (28.0, 572622), (29.0, 578394), (30.0, 587392), (31.0, 598217), (32.0, 609751), (33.0, 619658), (34.0, 626534), (35.0, 632082), (36.0, 636790), (37.0, 641169), (38.0, 645560), (39.0, 649759), (40.0, 653786), (41.0, 657686), (42.0, 661502), (43.0, 665246), (44.0, 668928), (45.0, 672554), (46.0, 676147), (47.0, 679730), (48.0, 683305), (49.0, 686870), (50.0, 690425), (51.0, 693975), (52.0, 697524), (53.0, 701056), (54.0, 704547), (55.0, 707970), (56.0, 711305), (57.0, 714535), (58.0, 717646), (59.0, 720613), (60.0, 723415), (61.0, 726032), (62.0, 728442), (63.0, 730629), (64.0, 732578), (65.0, 734280), (66.0, 735730), (67.0, 736928), (68.0, 737887), (69.0, 738627), (70.0, 739172), (71.0, 739540), (72.0, 739743), (73.0, 739792), (74.0, 739697), (75.0, 739472), (76.0, 739123), (77.0, 738658), (78.0, 738083), (79.0, 737406), (80.0, 737406)	(1.16)
World population = GRAPH(time/12) (Divided by 12 since these are annual data; unit: persons) (0.00, 3e+09), (1.00, 3.1e+09), (2.00, 3.1e+09), (3.00, 3.2e+09), (4.00, 3.3e+09), (5.00, 3.3e+09), (6.00, 3.4e+09), (7.00, 3.5e+09), (8.00, 3.6e+09), (9.00, 3.6e+09), (10.0, 3.7e+09), (11.0, 3.8e+09), (12.0, 3.9e+09), (13.0, 3.9e+09), (14.0, 4e+09), (15.0, 4.1e+09), (16.0, 4.2e+09), (17.0, 4.2e+09), (18.0, 4.3e+09), (19.0, 4.4e+09), (20.0, 4.5e+09), (21.0, 4.5e+09), (22.0, 4.6e+09), (23.0, 4.7e+09), (24.0, 4.8e+09), (25.0, 4.9e+09), (26.0, 4.9e+09), (27.0, 5e+09), (28.0, 5.1e+09), (29.0, 5.2e+09), (30.0, 5.3e+09), (31.0, 5.4e+09), (32.0, 5.4e+09), (33.0, 5.5e+09), (34.0, 5.6e+09), (35.0, 5.7e+09), (36.0, 5.8e+09), (37.0, 5.8e+09), (38.0, 5.9e+09), (39.0, 6e+09), (40.0, 6.1e+09), (41.0, 6.2e+09), (42.0, 6.2e+09), (43.0, 6.3e+09), (44.0, 6.4e+09), (45.0, 6.5e+09), (46.0, 6.5e+09), (47.0, 6.6e+09), (48.0, 6.7e+09), (49.0, 6.8e+09), (50.0, 6.8e+09), (51.0, 6.9e+09), (52.0, 7e+09), (53.0, 7e+09), (54.0, 7.1e+09), (55.0, 7.2e+09), (56.0, 7.2e+09), (57.0, 7.3e+09), (58.0, 7.4e+09), (59.0, 7.5e+09), (60.0, 7.5e+09), (61.0, 7.6e+09), (62.0, 7.6e+09), (63.0, 7.7e+09), (64.0, 7.8e+09), (65.0, 7.8e+09), (66.0, 7.9e+09), (67.0, 8e+09), (68.0, 8e+09), (69.0, 8.1e+09), (70.0, 8.1e+09), (71.0, 8.2e+09), (72.0, 8.3e+09), (73.0, 8.3e+09), (74.0, 8.4e+09), (75.0, 8.4e+09), (76.0, 8.5e+09), (77.0, 8.5e+09), (78.0, 8.6e+09), (79.0, 8.6e+09), (80.0, 8.7e+09)	(1.17)

High tourism growth implies accommodating overbooked hotel reservations for tourists who actually show up. Free cruises erode Cyprus' hotel EBITDA. The alternative is, however, angry tourists going off in hotel lobbies. Tourists have gotten angry at hotels before, but hotels have made the problem worse in recent years worldwide [16]. They have tightened check-in rules, doubled their renovations and increased the rate of overbooking by about 30 percent. The results can be explosive if one adds the record flight delays that travelers endure. Anyhow, free cruises to nearby Egypt and Israel sound much better than simply training employ-



Scenario-Driven Planning with System Dynamics, Figure 5

Four computed scenarios show how tourism growth might affect Cyprus' hotel EBITDA (smooth hE) and the island's environment, with carbon-dioxide (CO_2) pollution (adapted from and extending [30])

ees to handle unhappy guests that scream in hotel lobbies.

Eventually, as Cyprus' bed capacity increases and thereby catches up with tourism demand, there will be less overbooking and a few free cruises to erode Cyprus' hotel EBITDA. Given enough time for an initial bed capacity disequilibrium adjustment, in the long term, high tourism growth increases both hotel EBITDA (Fig. 5a) and cash [30].

In addition to their profound consequences for its hotel value-chain participants, Cyprus' tourism growth might also determine the fate of the island's environment. Depending on the island's population and emissions per person, high tourism growth implies high anthropogenic emissions feeding Cyprus' CO_2 Pollution. Anthropogenic CO_2 emissions attributed to the upward and downward movements of recurring tourist arrivals create much more stress and strain for the island's natural environment than a consistent stream of tourism with low seasonality would. High tourism growth lowers Cyprus' environmental footprint (EF). The summer 1974 Turkish *military intervention* has had a drastic negative effect on Cyprus' relative attractiveness because it reduced the island's free land by 41 percent.

Although qualitatively similar to the world's average EF after the invasion, Cyprus' environmental footprint is lower than the world's average EF (Fig. 5c), rendering the island's free area relatively less attractive as more foreign tourists visit. Manifested in the EF ratio (Fig. 5d), Cyprus becomes relatively less attractive as more visitors choose to vacation on the island's free area.

Qualitatively, Cyprus' CO_2 pollution scenarios (Fig. 5b) look exactly like the A2 scenario family of harmonized anthropogenic CO_2 emissions, which the *Intergovernmental Panel on Climate Change* (IPCC) computed to access the risks of human-induced climate change [54]. Like in the rest of the world, unless drastic changes in policy or technology alter the emissions per person ratio in the next 40 years, CO_2 pollution is expected to grow proportionally with Cyprus' tourism, degrading the island's environment.

Case 2: A Japanese Chemicals Keiretsu (JCK)

Home of NASA's *Johnson Space Center*, the Clear Lake region in Texas boasts strong high technology, biotechnology and specialty chemicals firms. Among them is JCK, whose recent investment helps the Clear Lake region continue its stalwart role in Houston's regional economic expansion [40].

An active member of a famous Japanese giant conglomerate, JCK's history begun in the late 1800s. Despite its long history, however, it has not been easy for JCK to evade the feedback loop that drives Japanese firms to manufacture outside Japan. Since the 1950s, with Japan still recovering from WWII, the better Japanese companies performed, the better their national currency did. But the better Japan's currency did, the harder it became for its firms to export. The higher the yen, the more expensive and, therefore, less competitive Japan's exports become. This simple loop explains JCK's manufacturing lineage from Japan to USA [34].

But the transition process behind this lineage is not that simple. JCK's use of SdP with SD reveals a lot about its strategy design and implementation tactics. The model below shows a tiny fragment of JCK's gigantic effort to reperceive itself. The firm wants to see its keiretsu transform into an agile, virtual enterprise network (VEN) of active agents that collaborate to achieve its transnational business goals. Although still flying low under the media's collective radar screen, VENs receive increased attention by strategic managers [29].

Sterman (see Chap. 17 and 18 in [75]) presents a generic value-chain management structure that can unearth what VENs are about. By becoming a VEN, JCK is poised to bring the necessary people and production processes together to form *autopoietic*, i. e., self-organizing, customer-centric value chains in the specialty chemicals industry. JCK decided to build its own plant in USA because the net present value (NPV) of the anticipated combined cash flow resulting from a merger with other specialty chemicals manufacturers in USA would have been less then the sum of the NPVs of the projected cash flows of the firms acting independently. Moreover, JCK's own technology transfer cost is so low that the internalization cost associated with a merger would far exceed supplier charges plus market transaction costs. To remain competitive [62], JCK will not integrate the activity but offshoot it as a branch of its VEN-becoming keiretsu. The plant will be fully operational in January 2008. In order to maximize the combined *n*et *p*resent *v*alue of *e*arnings *b*efore *i*nterest, *t*axes, *d*epreciation and *a*mortization, i. e., NPV(EBITDA), of its new USA plant and the existing one in Asia, JCK wishes to improve its USA sales revenue before production starts in USA.

JCK's pre-production marketing tactics aim at building a sales force to increase sales in USA. Until the completion of the new plant (Dec. 2007), JCK will keep importing chemicals from its plant in Asia. Once production starts in USA (Jan. 2008), then the flow of goods from Asia to USA stops, the plant in USA supplies the USA market and the flow of goods from USA to Asia begins.

Strategic scenarios are not new to the chemical industry [82]. SdP with SD helps this specialty chemicals producer integrate its business intelligence efforts with strategy design in anticipation of environmental change. Modeling JCK's strategic situation requires a comprehensive inquiry into the environmental causalities and equivocalities that dictate its actions. Computed strategic and tactical scenarios probe the combined consequences of environmental trends, changes in JCK's own strategy, as well as the moves of its current and future competitors. The section below describes briefly how JCK plans to implement its transnational strategy of balanced marketing and production. This takes the form of a system dynamics simulation model, which precedes the interpretation of its computed scenarios.

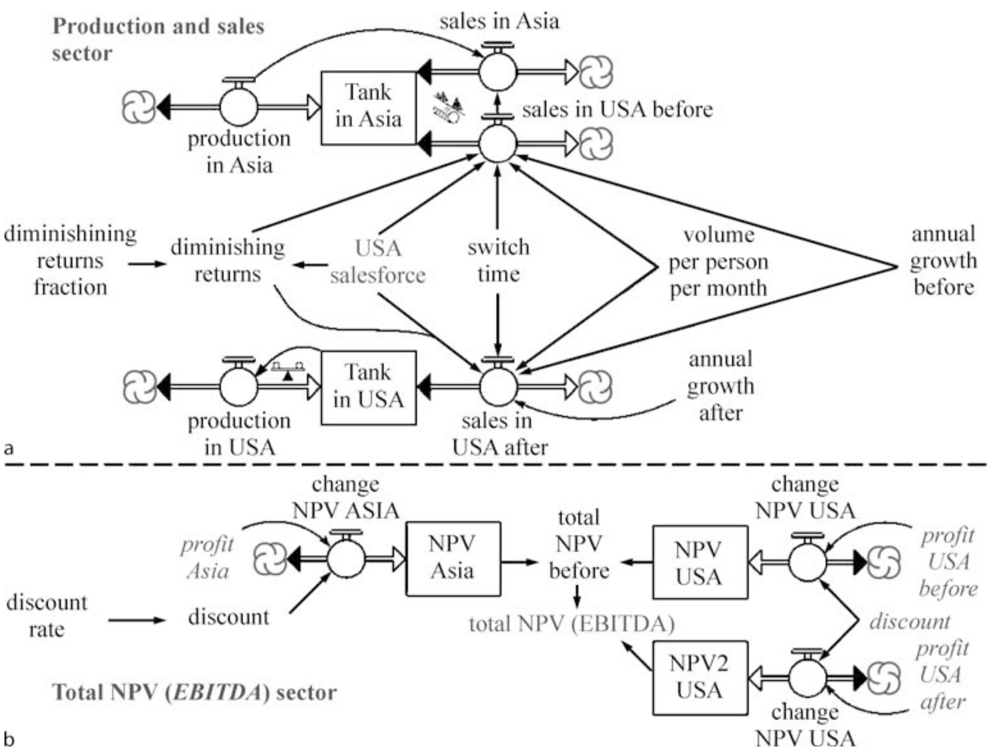
Model Description (Case 2)

The entire model has multiple sectors, four of which compute financial accounting data. Figure 6a shows the production and sales, and Fig. 6b the total NPV(EBITDA) model sectors. The corresponding algebra is in [34]. While JCK is building its USA factory, its factory in Asia makes and sells all specialty chemicals the USA market cannot yet absorb. This is what the *feed-forward* link from the production in Asia flow to the sales in Asia rate shows. The surplus demand JCK faces in Asia for its fine chemicals accounts for this rather unorthodox model structure. The surplus demand in Asia is the model's enabling *safety valve*, i. e., a major strategic assumption that renders tactical implementation feasible.

With the plant in Asia producing at full capacity before the switch, sales in the USA both depletes the tank in Asian stock and reduces sales in Asia. USA sales depend on JCK's USA sales force. But the size of this decision variable is just one determinant of sales in USA.

Sales productivity depends on many parameters, such as the annual growth before the switch rate of specialty chemicals in USA, the average expected volume a salesperson can sell per month as well as on the diminishing returns that sales people experience after the successful calls they initially make on their industrial customers. B2B or business to business, i. e., industrial marketing, can sometimes be as tough as B2C or business to customer, i. e., selling retail.

Time $t = 30$ months corresponds to January 2008, when the switch time converter cuts off the supply of JCK's chemicals from its plant in Asia. Ready by December 2007, the factory in the USA can supply the entire customer base its USA sales force will have been building for 30 months. As production in the USA begins, the sales in the USA before flow stops draining the tank in Asia and sales in Asia resume to match JCK's surplus demand there. Acting both as a production flow and as a continuous-review



Scenario-Driven Planning with System Dynamics, Figure 6
JCK's a production and sales, and b total NPV(EBITDA) model sectors (adapted from [34]; NPV = net present value, and EBITDA = earnings before interest, taxes, depreciation and amortization)

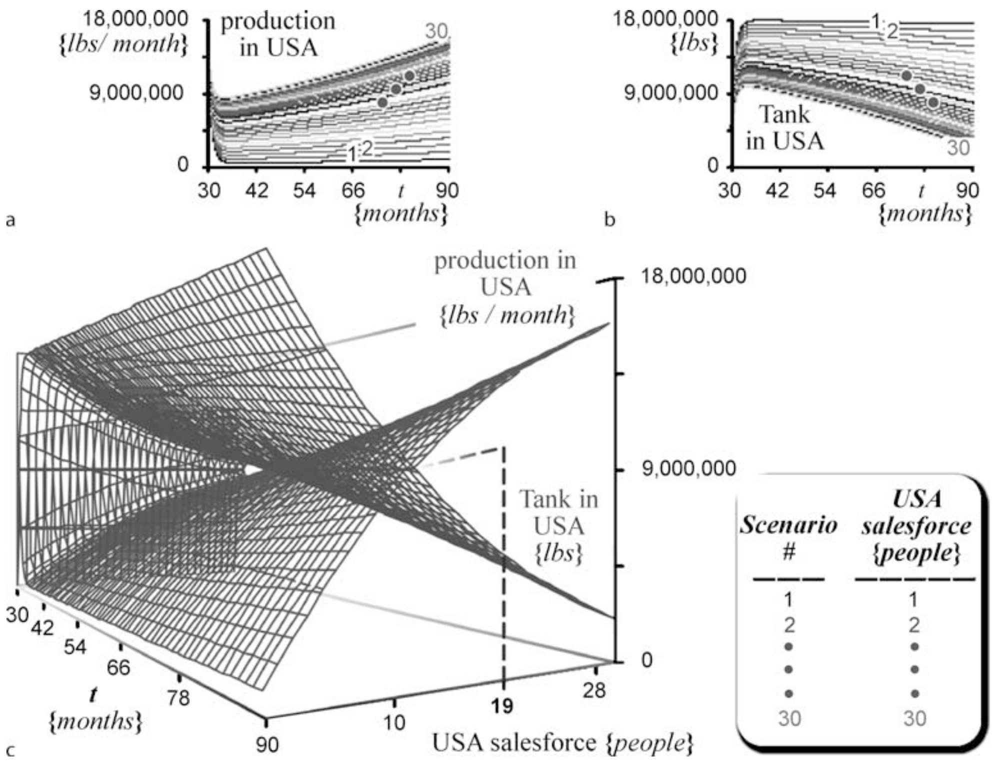
inventory order point, after January 2008, production in USA feeds the tank in USA stock of the rudimentary value-chain management structure on Fig. 6a.

Value chains entail stock and flow structures for the acquisition, storage and conversion of inputs into outputs, and the decision rules that govern the flows. The jet ski value chain includes, for example, hulls and bows that travel down monorail assembly paths. At each stage in the process, a stock of parts buffers production. This includes the inventory of fiberglass laminate between hull and bow acquisition and usage, the inventory of hulls and bows for the jet ski lower and upper structures, and the inventory of jet skis between dealer acquisition and sales. The decision rules governing the flows entail policies for ordering fiberglass laminate from suppliers, scheduling the spraying of preformed molds with layers of fiberglass laminate before assembly, shipping new jet skis to dealers and customer demand.

A typical firm's or VEN's value chain consists of cascades of supply chains, which often extend beyond a firm's boundaries. Effective value chain models must incorporate different agents and firms, including suppliers, the firm, distribution channels and customers. Scenario-driven planning with system dynamics is well suited for value chain modeling and policy design because value chains involve multiple stock and flow chains, with time lags or delays, and the decision rules governing the flows create multiple feedback loops among VEN members or value- and supply-chain partners (see Chap. 17 and 18 in [75]).

Back to JCK, its tank in the USA feeds information about its level back to production in the USA. Acting first as a decision point, production in the USA compares the tank in the USA level to the tank's capacity. If the tank is not full, then production in the USA places an order to itself and, once the USA factory has the requisite capacity, production in the USA refills the tank in the USA, but only until sales in the USA after the switch drains the tank. Then the cycle begins all over again.

Meanwhile, the profit in Asia, profit in the USA before and profit in the USA after sectors [34] perform all the financial accounting necessary to keep track of the transactions that take place in the value chain production and sales sector (Fig. 6a). As each scenario runs, the profit in Asia, the USA before and the USA after sectors feed the corresponding change in net present value (NPV) flows of



Scenario-Driven Planning with System Dynamics, Figure 7
Thirty computed scenarios show JCK's dual, smooth-switch and profitable purpose in production (adapted from [34])

the model's total NPV(EBITDA) sector (Fig. 6b). By adjusting each profit sector's EBITDA according to the discount rate, the change in NPV flows compute the total NPV(EBITDA) both in Asia and in the USA, both before and after JCK's January 2008 supply switch.

JCK's Computed Scenarios (Case 2)

Recall that the SdP with SD modeling-process spiral enabled our modeling team to crystallize JCK's strategic situation into the cyclical pattern that Fig. 3a shows. Although heavily disguised, the JCK measurement data and econometric sales functions let the system dynamics model compute scenarios to answer that razor-sharp optimization question the JCK executives asked:

> What size a USA sales force must we build in order to get a smooth switch in both sales and production in January 2008, and also to maximize the combined NPV(EBITDA) at our two plants in Asia and USA from now through 2012?

Treating the USA sales force policy parameter in the 'Sensi Specs…' menu item of *iThink*® allowed computing a set of 30 strategic scenarios. The 30 scenarios correspond to JCK's hiring from one to 30 sales people, respectively, to sell specialty chemicals to manufacturing firms in the USA, both before and after the January 2008 switch. Figures 7 and 8 show the 30 computed scenarios.

Figure 7c shows the response surfaces the production in USA rate and tank in the USA stock form after January 2008, in response to the 30 computed scenarios. The computed scenario that corresponds to JCK's building a USA sales force of 19 people achieves a smooth balance between sales in Asia and in the USA. Under this scenario, after January 2008, on the line where the two surfaces cross each other, not only the number of pounds of chemicals made and sold in Asia equals the number of pounds of chemicals made and sold in the USA, but as Fig. 7c shows, production in the USA also equals the tank in USA stock. So hiring 19 sales people now meets JCK's smooth switch in sales and production objective. But how?

How does producing and selling in the USA at rates equal to the corresponding rates in Asia constitute a fair response to JCK's smooth switch objective? The JCK executives seemed to accept this at face value. But our team had to clearly explain the dynamics of JCK's rudimentary USA value chain (Fig. 6a), in order to unearth what the USA member of this transnational VEN-becoming keiretsu might be up to.

It looks simple, but the value chain of the production and sales sector on Fig. 6a can show the same amplification symptoms that much more elaborate value chains show when they fall pray to bullwhip effects. Locally rational policies that create smooth and stable adjustment of individual business units can, through their interaction with other functions and firms, cause oscillation and instability. Figure 8a shows the profound consequences of JCK's switch for its value chain in the USA. Because of the sudden switch in January 2008, the computed scenarios cause 30 sudden step changes. Both variables' adjustment rates increase, but the tank in the USA stock's amplification remains almost constant below 50 percent. As customer demand steps up, so do both metrics' new equilibrium points, but in direct proportion to the step increase in customer demand in the USA.

The 30 computed scenarios confirm Sterman's argument that, while the magnitude of amplification depends on stock adjustment times and delivery lags, its existence does not. No matter how drastically customers and firms downstream in a value chain change an orders' magnitude, they cannot affect supply chain amplification. Value chain managers must never blame customers and downstream firms or their forecasts for bullwhip effects. The production in USA amplification is almost double the tank in USA for a small USA sales force, suggesting that JCK's USA factory faces much larger changes in demand than its sales people do. Although temporary, during its disequilibrium adjustment, the tank in the USA consistently overshoots the new equilibrium points that it seeks after the switch (Fig. 7b), an inevitable consequence of stock and flow structure. Customers are innocent, but JCK's value chain structure is not:

First, the tank in the USA stock adjustment process creates significant amplification of the production in the USA rate. Though the tank in the USA relative amplification is 36.18 percent under the USA sales force = 1 scenario, for example, the relative amplification of production in the USA (Fig. 8a) increases by a maximum of more than 90 percent (the peak production in the USA rate, after t = 30 months, divided by the minimum production in the USA rate $= 11,766,430.01/1,026,107.64 = 91.28$ percent). The *amplification ratio*, i. e., the ratio of maximum change in output to maximum change in input, therefore is $91.28\%/36.18\% = 2.52$. A one-percent increase in demand for JCK's chemicals causes a 2.52 percent surge in demand at JCK's USA plant. While the amplification ratio magnitude depends on the stock adjustment times and delivery lags, its existence does not (see p. 673 in [75]).

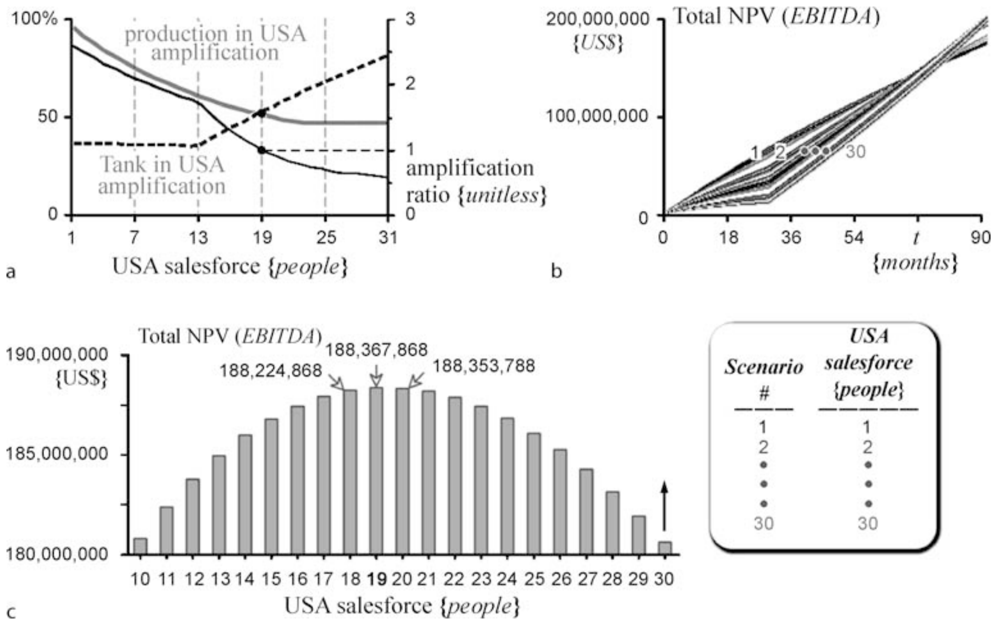
Second, amplification is temporary. In the long run, a one-percent increase in sales in the USA after leads to a one-percent increase in production in the USA. After two-adjustment times, i. e., two months, production in the USA gradually drops (Fig. 7a). During the disequilibrium adjustment, however, production in the USA overshoots its new equilibrium, an inevitable consequence of the stock and flow structure of customer-supplier value chains, no matter how tiny or simple they are. The only way the tank in the USA stock can increase is for its inflow production in the USA rate (order rate) to exceed its outflow rate sales in the USA after (Fig. 6a). Within a VEN's or keiretsu's customer-supplier value chain, supply agents face much larger changes in demand than finished-goods inventory, and the surge in demand is temporary.

The computed scenarios show that as the USA sales force increases, the production in the USA's rate of amplification declines because its new long-term equilibrium point is closer to its initial jump in January 2008. Conversely, as the tank in the USA stock's long-term equilibrium point remains consistently high because of the larger USA sales force, its relative amplification begins to rise. Since the two variables' relative amplification moves in opposite directions, eventually, they meet. What a coincidence! They meet above the USA sales force = 19 people. Now, is this not a much better interpretation of the word 'smooth' in fair response to JCK's smooth-switch performance purpose? The answer to JCK is now pertinent to its balancing its value chain in USA. With a USA sales force = 19, JCK's value chain components show equal relative amplification to sudden changes in demand, attaining nothing less than a magnificent amplification ratio = 1. Now that is smooth!

But what of profitability? JCK's polite executives said: "maximize… combined… NPV". In the time domain (Fig. 8b), total NPV(EBITDA) creates intricate dynamics that obscures the USA sales force effect. But the phase plot on Fig. 8c clearly shows a concave down behavior along the USA sales force: USA sales force = 19 maximizes the two plants' combined total NPV(EBITDA).

Future Directions

The above cases show how scenario-driven planning with system dynamics helps control performance by enabling organizational learning and the management of uncertainty. The strategic intelligence system that SdP with SD provides rests on the idea of a collective inquiry, which translates the environmental 'macrocosm' and a firm's



Scenario-Driven Planning with System Dynamics, Figure 8
Thirty computed scenarios show how hiring a sales force of 19 people in the USA might maximize JCK's NPV(EBITDA), and thereby fulfill its dual, smooth-switch and profitable purpose (adapted from [34])

'microcosm' into a shared causal map with computed scenarios. Informed discussion then takes place. Seeing SdP with SD as an inquiry system might help the outcomes of the situation formulation-solution-implementation sequence, each stage built on successive learning.

Strategic situations are complex and uncertain. Because planning is directed toward the future, predictions of changes in the environment are indispensable components of it. Conventional forecasting by itself provides no cohesive way of understanding the effect of changes that might occur in the future. Conversely, SdP with SD and its computed scenarios provide strategic intelligence and a link from traditional forecasting to modern interactive planning systems. In today's quest for managers who are more leaders than conciliators, the strategists' or executives' interest in scenarios must be welcomed. A clearer delineation of SdP with SD might make it a very rich field of application and research.

The SdP with SD inquiry system on Fig. 2 includes several contributions. *First*, by translating the environmental *macrocosm* and the firm *microcosm* into a common context for conceptualization, the requisites of theory building can be addressed. Planning analysts no longer have to operate piecemeal. A theory and a dominant logic typically emerge from shared perceptions about a firm, its environment and stakeholder purposes through model construction.

Second, the outputs of the strategic management process activities build on each other as successive layers. The SAST loop on Fig. 2 follows the counterclockwise direction of multiperspective dialectics [47]. This process allows adjustment of individual and organizational theories and logic, leading to an evolutionary interpretation of the real system that strategic decisions target.

Third, the inquiry system of Fig. 2 enables flexible support for all phases of strategy design. Problem finding or forming, or situation formulation receives equal attention as problem solving.

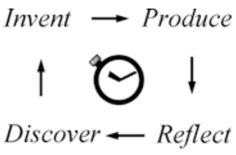
SdP with SD helps open up the black box of decision makers' mental models, so they can specify the ideas and rules they apply. That in turn helps enrich their language and label system, organizational capability and knowledge, and strategic decision processing system. Computed scenarios bring about transformation rules not previously thought of as well as new variables and interaction paths.

As an entity, each decision maker has a local scope and deals only with specific variables and access paths to other entities. But success factors are not etched in stone. Often, we only observe a representative state of each entity, namely, locally meaningful variables and parts of a scenario. This representativeness changes dynamically in the process of computing scenarios. Beyond the purely technical advantages of computed scenarios, planning becomes interactive, and language and label systems render themselves more adequate, effective and precise. Their associated organizational capability develops even more. In addition, the minor and major assumptions in decision makers' mental models surface as computed scenarios specify the conditions under which performance changes.

A line of great immediate concern requires researchers and practitioners alike to explore the modeling process behind SdP with SD. For the sake of realism, to make negotiated perceptions of reality explicit, we need representations where strategic real options and self-interest projections mold the way in which managers incorporate their observations and interpretations into strategy models. This is an unavoidable, most challenging path to tread, if we want to build a dialectical debate into the strategy design process.

Do we really want to? Yes because:

1. The traditional hierarchical organization dogma has been planning, managing and controlling, whereas the new reality of the learning organization incorporates vision, values and mental models. It entails training managers and teams in the IPRD learning cycle conceived by Dewey [14] (cf. Senge and Sterman [71]):



2. In the strategic management process (SMP) evolution, planning is evolving too, from objective-driven to budget-driven to strategy-driven to scenario-driven planning with system dynamics (SdP with SD, see pp. 271–272 in [32]).
3. The inquiry system that mediates the restructuring of organizational *theory in use* [68] determines the quality of organizational learning.

By looking into the dynamics of strategy design and the resulting performance of firms, the SdP with SD framework on Fig. 2 might let managers, planners and business researchers see the tremendous potential of computed strategic scenarios. They might choose to build intelligence systems around SdP with SD to create insight for strategy design. They will be building real knowledge in the process, while developing capability for institutional learning. Both Pascale [59] and de Geus [13] see the capability to speed up institutional learning as a truly sustainable competitive advantage.

Bibliography

Primary Literature

1. Acar W (1983) Toward a theory of problem formulation and the planning of change: Causal mapping and dialectical debate in situation formulation. UMI, Ann Arbor
2. Ackoff RL (1981) Creating the corporate future. Wiley, New York
3. Ackoff RL, Emery FE (1972) On purposeful systems. Aldine-Atherton, Chicago
4. Amara R, Lipinski AJ (1983) Business planning for an uncertain future: Scenarios and strategies. Pergamon, New York
5. Anderson TJ (2000) Real options analysis in strategic decision making: An applied approach in a dual options framework. J Appl Manag Stud 9(2):235–255
6. Ansoff HI (1985) Conceptual underpinnings of systematic strategic management. Eur J Oper Res 19(1):2–19
7. Ansoff HI, McDonnell E (1990) Implanting strategic management, 2nd edn. Prentice-Hall, New York
8. Brauers J, Weber M (1988) A new method of scenario analysis for strategic planning. J Forecast 7(1):31–47
9. Brenkert AL (1998) Carbon dioxide emission estimates from fossil-fuel burning, hydraulic cement production, and gas flaring for 1995 on a one degree grid cell basis. Oak Ridge National Laboratory, Carbon Dioxide Information Analysis Center, Oak Ridge, TN (Database: NDP-058A 2-1998)
10. Christensen CM (1997) The innovator's dilemma: When new technologies cause great firms to fail. Harvard Business School Press, Cambridge
11. CYSTAT (2000) Cyprus key figures: Tourism. The Statistical Service of Cyprus (CYSTAT), Nicosia
12. Daft RL, Weick KE (1984) Toward a model of organizations as interpretation systems. Acad Manag Rev 9:284–295
13. de Geus AP (1992) Modelling to predict or to learn? Eur J Oper Res 59(1):1–5
14. Dewey J (1938) Logic: The theory of inquiry. Holt, Rinehart and Winston, New York
15. Donaldson L (1992) The Weick stuff: Managing beyond games. Organ Sci 3(4):461–466
16. Drucker J, Higgins M (2001) Hotel rage: Losing it in the lobby. Wall Street J (Fri 16 Feb):W1–W7
17. Duncan RB (1972) Characteristics of organizational environments and perceived environmental uncertainty. Adm Sci Q 17:313–327
18. Eberlein RL (2002) Vensim® PLE Software, V 5.2a. Ventana Systems Inc, Harvard
19. Eden C (1994) Cognitive mapping and problem structuring for system dynamics model building. Syst Dyn Rev 10(3):257–276
20. Eden C (2004) Analyzing cognitive maps to help structure issues or problems. Eur J Oper Res 159:673–686
21. Eilon S (1984) The Art of Reckoning: Analysis of performance criteria. Academic Press, London
22. Emery FE, Trist EL (1965) The causal texture of organizational environments. Hum Relat 18:21–32
23. Forrester JW (1958) Industrial dynamics: A major breakthrough for decision makers. Harvard Bus Rev 36(4):37–66
24. Forrester JW (1961) Industrial dynamics. MIT Press, Cambridge
25. Forrester JW (1987) Lessons from system dynamics modeling. Syst Dyn Rev 3(2):136–149

26. Forrester JW (1992) Policies, decisions and information sources for modeling. Eur J Oper Res 59(1):42–63
27. Georgantzas NC (1995) Strategy design tradeoffs-free. Hum Syst Manag 14(2):149–161
28. Georgantzas NC (2001) Simulation modeling. In: Warner M (ed) International encyclopedia of business and management, 2nd edn. Thomson Learning, London, pp 5861–5872
29. Georgantzas NC (2001) Virtual enterprise networks: The fifth element of corporate governance. Hum Syst Manag 20(3):171–188, with a 2003 reprint ICFAI J Corp Gov 2(4):67–91
30. Georgantzas NC (2003) Tourism dynamics: Cyprus' hotel value chain and profitability. Syst Dyn Rev 19(3):175–212
31. Georgantzas NC (2007) Digest® wisdom: Collaborate for win-win human systems. In: Shi Y et al (eds) Advances in multiple criteria decision making and human systems management. IOS Press, Amsterdam, pp 341–371
32. Georgantzas NC, Acar W (1995) Scenario-driven planning: Learning to manage strategic uncertainty. Greenwood-Quorum, Westport
33. Georgantzas NC, Katsamakas E (2007) Disruptive innovation strategy effects on hard-disk maker population: A system dynamics study. Inf Resour Manag J 20(2):90–107
34. Georgantzas NC, Sasai K, Schrömbgens P, Richtenburg K et al (2002) A chemical firm's penetration strategy, balance and profitability. In: Proc of the 20th international system dynamics society conference, 28 Jul–1 Aug, Villa Igiea, Palermo, Italy
35. Gharajedaghi J (1999) Systems thinking: Managing chaos and complexity – A platform for designing business architecture. Butterworth-Heinemann, Boston
36. Godet M (1987) Scenarios and strategic management: Prospective et planification stratégique. Butterworths, London
37. Godet M, Roubelat F (1996) Creating the future: The use and misuse of scenarios. Long Range Plan 29(2):164–171
38. Hax AC, Majluf NS (1996) The strategy concept and process: A pragmatic approach, 2nd edn. Prentice Hall, Upper Saddle River
39. Helmer O (1983) Looking forward. Sage, Beverly Hills
40. Hodgin RF (2001) Clear Lake Area Industry and Projections 2001. Center for Economic Development and Research, University of Houston-Clear Lake, Houston
41. Huss WR, Honton EJ (1987) Scenario planning: What style should you use? Long Range Plan 20(4):21–29
42. Istvan RL (1992) A new productivity paradigm for competitive advantage. Strateg Manag J 13(7):525–537
43. Jarillo JC (1988) On strategic networks. Strateg Manag J 9(1):31–41
44. Jarillo JC, Martínez JI (1990) Different roles for subsidiaries: The case of multinational corporations in Spain. Manag J 11(7):501–512
45. Kahn H, Wiener AJ (1967) The next thirty-three years: A framework for speculation. Daedalus 96(3):705–732
46. Lissack MR, Roos J (1999) The next common sense: The e-manager's guide to mastering complexity. Nicholas Brealey Publishing, London
47. Mason RO, Mitroff II (1981) Challenging strategic planning assumptions. Wiley, New York
48. Miller D, Friesen PH (1983) Strategy making and environment: The third link. Strateg Manag J 4:221–235
49. Millet SM, Randles F (1986) Scenarios for strategic business planning: A case history for aerospace and defence companies. Interfaces. 16(6):64–72
50. Mojtahedzadeh MT, Andersen D, Richardson GP (2004) Using Digest® to implement the pathway participation method for detecting influential system structure. Syst Dyn Rev 20(1):1–20
51. Moore GA (1991) Crossing the chasm. Harper-Collins, New York
52. Morecroft JDW (1985) Rationality in the analysis of behavioral simulation models. Manag Sci 31:900–916
53. Morecroft JDW (1988) System dynamics and microworlds for policymakers. Eur J Oper Res 35:310–320
54. Nakicenovic N, Davidson O, Davis G et al (2000) Summary for policymakers–Emissions scenarios: A special report of working group III of the intergovernmental panel on climate change (IPCC). World Meteorological Organization (WMO) and United Nations Environment Programme (UNEP), New York
55. Nicolis JS (1986) Dynamics of hierarchical systems: An evolutionary approach. Springer, Berlin
56. Oliva R (2004) Model structure analysis through graph theory: Partition heuristics and feedback structure decomposition. Syst Dyn Rev 20(4):313–336
57. Oliva R, Mojtahedzadeh MT (2004) Keep it simple: A dominance assessment of short feedback loops. In: Proc of the 22nd international system dynamics society conference, 25–29 July 2004, Keble College, Oxford University, Oxford UK
58. Ozbekhan H (1977) The future of Paris: A systems study in strategic urban planning. Philos Trans Royal Soc London. 387:523–544
59. Pascale RT (1984) Perspectives on strategy: The real story behind Honda's success. California Manag Rev 26(Spring):47–72
60. Pine BJ-I, Victor B, Boynton AC (1993) Making mass customization work. Harvard Bus Rev 71(5):108–115
61. Porter ME (1985) Competitive advantage: Creating and sustaining superior performance. Free Press, New York
62. Porter ME (1991) Towards a dynamic theory of strategy. Strateg Manag J 12(Winter Special Issue):95–117
63. Randers J (1980) Guidelines for model conceptualization. In: Randers J (ed) Elements of the system dynamics method. MIT Press, Cambridge, pp 117–139
64. Raynor ME (2007) The strategy paradox: Why committing to success leads to failure [And What to Do About It]. Currency-Doubleday, New York
65. Repenning NP (2003) Selling system dynamics to (other) social scientists. Syst Dyn Rev 19(4):303–327
66. Richardson GP (1991) Feedback thought in social science and systems theory. University of Pennsylvania Press, Philadelphia
67. Richmond B et al (2006) iThink® Software V 9.0.2. iSee Systems™, Lebanon
68. Schön D (1983) Organizational learning. In: Morgan G (ed) Beyond method. Sage, London
69. Schrage M (1991) Spreadsheets: Bulking up on data. San Francisco Examiner
70. Schwenk CR (1984) Cognitive simplification processes in strategic decision making. Strateg Manag J 5(2):111–128
71. Senge PM, Sterman JD (1992) Systems thinking and organizational learning: Acting locally and thinking globally in the organization of the future. Eur J Oper Res 59(1):137–150
72. Simon HA (1979) Rational decision making in business organizations. Am Econ Rev 69(4):497–509
73. Singer AE (1992) Strategy as rationality. Hum Syst Manag 11(1):7–21
74. Singer AE (1994) Strategy as moral philosophy. Strateg Manag J 15:191–213

75. Sterman JD (2000) Business dynamics: Systems thinking and modeling for a complex world. Irwin McGraw-Hill, Boston
76. Sterman JD, Wittenberg J (1999) Path dependence, competition and succession in the dynamics of scientific revolution. Organ Sci 10(3, Special issue: Application of complexity theory to organization science, May–Jun):322–341
77. Turner F (1997) Foreword: Chaos and social science. In: Eve RA, Horsfall S, Lee ME (eds) Chaos, complexity and sociology. Sage, Thousand Oaks, pp xi–xxvii
78. Wack P (1985) Scenarios: Uncharted waters ahead. Harvard Bus Rev 63(5):73–89
79. Wack P (1985) Scenarios: Shooting the rapids. Harvard Bus Rev 63(6):139–150
80. Zeleny M (1988) Parallelism, integration, autocoordination and ambiguity in human support systems. In: Gupta MM, Yamakawa T (eds) Fuzzy logic in knowledge-based systems, decision and control. Elsevier Science, North Holland, pp 107–122
81. Zeleny M (2005) Human systems management: Integrating knowledge, management and systems. World Scientific, Hackensack
82. Zentner RD (1987) Scenarios and the chemical industry. Chem Marketing Manag (Spring):21–25

Books and Reviews

Bazerman MH, Watkins MD (2004) Predictable surprises: The disasters you should have seen coming and how to prevent them. Harvard Business School Press, Boston

Bower G, Morrow D (1990) Mental models in narrative comprehension. Science 247(4938):44–48

Mittelstaedt RE (2005) Will your next mistake be fatal? Avoiding the chain of mistakes that can destroy your organization. Wharton School Publishing, Upper Saddle River

Morecroft JDW (2007) Strategic modeling and business dynamics: A feedback systems approach. Wiley, West Sussex

Schnaars SP (1989) Megamistakes: Forecasting and the myth of rapid technological change. The Free Press, New York

Schwartz P (1991) The art of the long view. Doubleday-Currency, New York

Tuchman B (1985) The March of folly: From troy to vietnam. Ballantine Books, New York

Vennix JAM (1996) Group model building: Facilitating team learning using system dynamics. Wiley, Chichester

Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space

Gert Zöller[1], Sebastian Hainzl[2],
Yehuda Ben-Zion[3], Matthias Holschneider[1]

[1] Institute of Mathematics and Centre for Dynamics of Complex Systems, University of Potsdam, Potsdam, Germany

[2] GFZ German Research Centre for Geosciences, Potsdam, Germany

[3] Department of Earth Sciences, University of Southern California, Los Angeles, USA

Article Outline

Glossary

Bayesian analysis A model estimation technique that accounts for incomplete knowledge. Bayes' theorem is a mathematical formulation of how an a priori estimate of the probability of an event can be updated, if a new information becomes available.

Critical earthquake concept The occurrence of large earthquakes may be described in terms of statistical physics and thermodynamics. In this view, an earthquake can be interpreted as a critical phase transition in a system with many degrees of freedom. The preparatory process is characterized by acceleration of the seismic moment release and growth of the spatial correlation length as in the percolation model. This interpretation of earthquake occurrence is referred to as the critical earthquake process.

Earthquake forecast/prediction The forecast or prediction of an earthquake is a statement about time, hypocenter location, magnitude, and probability of occurrence of an individual future event within reasonable error ranges.

Fault model A fault model calculates the evolution of slip, stress, and related quantities on a fault segment or a fault region. The range of fault models varies from conceptual models of cellular automaton or slider-block type to detailed models for particular faults.

Probability A quantitative measure of the likelihood for an outcome of a random process. In the case of repeating a random experiment a large number of times (e. g. flipping a coin), the probability is the relative frequency of a possible outcome (e. g. head). A different view of probability is used in the → Bayesian analysis.

Seismic hazard The probability that a given magnitude (or peak ground acceleration) is exceeded in a seismic source zone within a pre-defined time interval, e. g. 50 years, is denoted as the seismic hazard.

Self-organized criticality
Self-organized criticality (SOC) as introduced by Bak [2] is the ability of a system to organize itself in

the vicinity of a critical point independently of values of physical parameters of the system and initial conditions. Self-organized critical systems are characterized by various power law distributions. Examples include models of sandpiles and forest-fires.

Definition of the Subject

The most fundamental question in earthquake science is whether earthquake prediction is possible. Related issues include the following: Can a prediction of earthquakes solely based on the emergence of seismicity patterns be reliable? In other words, is there a single or several "magic" parameters, which become anomalous prior to a large earthquake? Are pure observational methods without specific physical understanding, like the pattern recognition approach of Keilis–Borok and co-workers [41], sufficient? Taking into account that earthquakes are monitored continuously only for about 100 years and the best available data sets ("earthquake catalogs") cover only a few decades, it seems questionable to forecast earthquakes solely on the basis of observed seismicity patterns. This is because large earthquakes have recurrence periods of decades to centuries; consequently, data sets for most regions include less than ten large events making a reliable statistical testing questionable.

In the studies discussed here, the goal is not to forecast individual earthquakes. Instead, we aim at developing a combined approach based on numerical modeling and data analysis in order to understand seismicity and the emergence of patterns in the occurrence of earthquakes. The discussion and interpretation of seismcity in terms of statistical physics leads to the concept of "critical states", i. e. states in the seismic cycle with an increased probability for abrupt changes involving large earthquakes. A more general goal of this work is to provide perspectives for the understanding of the relevant mechanisms and to give outlines for developments related to time-dependent seismic hazard.

Introduction

Several empirical relationships for the occurrence of seismicity are well-known. The most common one is probably the Gutenberg–Richter law [30] for the relation between frequency and magnitude of earthquakes in a large seismically active region,

$$\log N = a - bM\,, \tag{1}$$

where N is the frequency of earthquakes with magnitude equal to or greater than M; a is a measure of the overall seismicity level in the region and the b value determines the relation between large and small earthquakes. The Gutenberg–Richter law provides an important constraint for the design of physical models and serves as a key ingredient for seismic hazard estimations. Statistical relations for the temporal occurrence of large events are less well known, because the corresponding data records are too short.

Several additional problems exist in the understanding and interpretation of observed seismicity patterns. First, it is important to decide whether an observed pattern has a physical origin or is an artifact, arising for example from inhomogeneous reporting or from man-made seismicity like quarry blasts or explosions [69]. Second, the non-artificial events have to be analyzed with respect to their underlying mechanisms. This leads to an inverse problem with a non-unique solution, which can be illustrated for the most pronounced observed temporal pattern associated with aftershocks. It is empirically known that the earthquake rate $\dot{N}$ after a large event at time t_{M} follows the Omori–Utsu law [49,67]

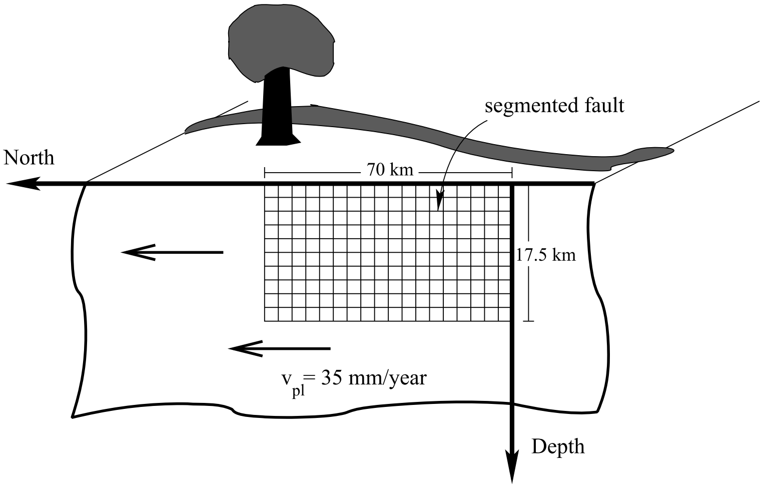
$$\dot{N} = \frac{K}{(c + t - t_{\mathrm{M}})^p}\,, \tag{2}$$

where t is the time, K and c are constants, and the Omori exponent p is close to unity. In particular, aftershocks are an almost universal phenomenon; that is, they are observed nearly after each mainshock. The underlying mechanisms leading to aftershocks are, however, unknown. Various physical models have been designed to explain aftershock occurrence following Eq. (2). These models include viscoelasticity [32], pore fluid flow [46], damage rheology [9,57], and rate-state friction [24]. The question which mechanism or combination of mechanisms is relevant in a given fault zone remains open. Detailed comparisons of observed and modeled seismicity with respect to the aftershock rate, the duration of aftershock sequences, the dependence on the mainshock size, and other features are necessary to address this problem. Additionally, results from lab experiments on rupture dynamics and satellite observations of deformation provide important information for the design of such models.

Apart from aftershock activity, other seismicity patterns are occasionally associated with observations, including foreshocks [39], seismic quiescence [34,72,78], and accelerating moment release [17,38]. These patterns have been documented in several cases before large earthquakes. They occur, however, far less frequently than aftershocks. For example, foreshocks are known to preceed only 20–30% of large earthquakes [71]. Therefore, their predictive power is questionable. Moreover, it is not clear whether these patterns can be attributed to physical processes or to random fluctuations in the highly sparse and noisy earthquake catalogs. This problem can be addressed by using fault models which simulate long and complete earthquake sequences over thousands of years. If the models capture the main features of the underlying physics, the occurrence of seismicity patterns can be studied with reasonable statistics. The main ingredients of such models are the geometry of a fault region, empirically known constitutive laws, spatial heterogeneities, and stress and displacement functions following dislocation theory [20,47]. In order to allow for detailed studies of the relations between the imposed mechanisms and the observed seismicity functions, it is important that the number of adjustable parameters is limited.

It is emphasized that these models do not aim to reproduce an observed earthquake catalog in detail. Instead, the main goal is to address questions like: Why is the Parkfield segment of the San Andreas fault characterized by relatively regular occurrence of earthquakes with magnitude $M \approx 6$, while on the San Jacinto fault in California the properties of earthquake occurrence are more irregular? Basic models for seismicity are mainly based on one or more solid blocks, which are driven by a plate over a sliding surface. The plate and the blocks are connected with springs. This model can generate stick-slip events considered to represent earthquakes. The slider-block models can produce a wide range of complexity, beginning with a single block model leading to periodic occurrence of events of uniform size, and progressing to an array of connected blocks [18] leading to complex sequences of events with variable size. In order to reduce the computational effort cellular automata are commonly used [42,48]. Mathematically, these models include maps instead of differential equations; physically, this corresponds to instantaneously occurring slip events, neglecting inertia effects. The main ingredients of slider-block and cellular automaton models are (1) external driving (plate motion), and (2) sudden local change of system parameters (stress), when a critical value (material strength) is reached, followed by an avalanche of block slips (stress drop and coseismic stress transfer during an earthquake). While the first process lasts for years to several hundred years, the second occurs on a time scale of a few seconds. The simplest model including these features has been formulated by Reid [52] and is known as *Reid's elastic rebound theory*; in terms of slider-block models, this corresponds to a single block model with constant plate velocity. Accounting for spatial heterogeneity and fault segmentation, many interacting blocks, or fault segments, have to be considered. This leads to a spatiotemporal stress field instead of a single stress value. In general, the material strength will also become space-dependent. Such a model framework can be treated with the methodology of statistical physics similar to the Ising model or percolation models [43]. In this context, large earthquakes are associated with second-order phase transitions [2,59,64]. The view of earthquakes as phase transitions in a system with many degrees of freedom and an underlying critical point, is hereinafter referred to as the "critical point concept". The period before such a phase transition is expected to be characterized by a preparation process including development of power laws and growing spatial correlation length [14]. However, depending on the parameters of a model, different situations are conceivable: the system trajectory can enter the critical state and the critical point frequently ("supercritical") or it may never becomes critical ("subcritical"). A case of special interest is the class of models [2] showing *self-organized criticality* (SOC), which have their origin in a simple cellular automaton model for a sandpile [3]. In this case the system drives itself permanently to the vicinity of the critical point with almost scale-free characteristics. Consequently, each small event can grow into a large earthquake with some probability [28].

Long simulations of earthquake activity can be used to calculate statistical features like the recurrence time distribution of large earthquakes and the frequency-size distribution with high precision. Despite the scaling behavior (Eq. (1)) in the earthquake magnitudes for small and intermediate earthquakes, which is observed for many sets of model parameters, clear deviations become visible for large magnitudes. Such deviations are known from real catalogs, but their statistical significance is not clear in all cases. The model simulations suggest that deviations from scaling for strong earthquakes can be attributed to physical properties. One important property is the spatial disorder of brittle parameters of the fault. The presence of strong heterogeneities suppresses system-wide events with some probability, whereas such events can evolve more easily on smooth faults. The degree of quenched (time-independent) spatial heterogeneity turns out to be a key parameter for statistical and dynamical properties of seismicity [5,12,80]. This includes the temporal regularity of mainshock occurrence, various properties of the stress and displacement fields, and a spontaneous mode-switching between different dynamical regimes without changing parameters. The degree of heterogeneity can act as a tuning parameter that allows for a continuous change of the model dynamics between the end-member cases of supercritical and subcritical behavior. Such a dependence, which is observed also for other parameters, can be visual-



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 1
Sketch of the fault model framework

ized in phase diagrams similar to the phase diagram for the different aggregate states of water [22,79,80]. For increasing complexity of a model, the number of axes of the phase diagram, representing the relevant model parameters, will increase. The above mentioned question of distinguishing different faults like the Parkfield segment and the San Jacinto fault can be rephrased as the problem of assigning the faults to different regions in such a diagram. An important step in this direction is the physical modeling of observed seismicity patterns, including universal patterns like aftershocks (Eq. (2)), common fluctuations like foreshocks and the acceleration of seismic energy release before large earthquakes. The latter phenomenon which sometimes occurs over large regions including more than one fault, can be interpreted in terms of the approach towards a critical point. This view is supported by an observational study of the growth of the spatial correlation length which is a different aspect of the same underlying physics [73,75,76,77].

The establishment of relationships between model parameters and observational features may be used to tune the model towards a specific fault zone, and use the tuned fault model for practical applications of seismic hazard estimations. Toward this end the recurrence time distribution of large earthquakes is needed. Since observational data records are often short and noisy, the use of Bayesian probability theory is helpful for the estimation of uncertain model parameters, and the incorporation of various types of observational data in seismic hazard estimations. The Parkfield segment, as one of the best monitored seismically-active regions, serves as an excellent natural laboratory for such a case study. A discussed example illustrates how partially known parameters like the stress drop and the seismic hazard can be estimated by combining numerical models and observational data [74].

In Section “Modeling Seismicity in Real Fault Regions”, the physical fault model used for the discussed studies is described. Results from numerical simulations are presented in Sect. “Results”. A summary is given in Sect. “Summary and Conclusions”.

Modeling Seismicity in Real Fault Regions

Numerous frameworks have been used to simulate seismicity (see e. g. [6,9,10,18,32,37,48] and references therein). These include slider-block models, cellular automata, “inherently discrete” fault models where the discreteness is an inherent feature of the imposed physics, and continuum models. In this section we illustrate how a fault model (Fig. 1) can be adjusted in order to simulate seismicity of a real fault region, e. g. the Parkfield segment of the San Andreas fault in California.

Fault Geometry and Model Framework

A first constraint for a specific model is to represent the geometry of the fault segment. As shown in Fig. 2, the region of Parkfield is characterized by a distribution of fault segments, which have generally the same orientation. It is therefore reasonable to map these segments in the model on a plane intersecting the surface at a straight line from SE to NW. The dimensions of the fault segment for modeling (Fig. 1) are chosen to be 70 km in length and 17.5 km in depth. As discussed in [10], this geometry corresponds approximately to the San Andreas fault near Parkfield. The entire fault is an infinite half-plane, but the brittle processes are calculated on the above rectangular section referred to below as the computational grid. The computational grid is discretized to 128 × 32 cells of uniform size, where stress and slip are calculated. The size of the cells is not related here to observations; rather it depends on the magnitude range under consideration and the computational effort. The failure of a single cell defines the lowest magnitude. A higher resolution of the grid with same overall dimensions increases the magnitude range, because the magnitude is calculated from the slip of all cells during an earthquake. Following Ben–Zion and Rice [10], the material surrounding the fault is assumed to be a homogeneous elastic half space, which is characterized by elastic parameters and a related Green's function:

1. The elastic properties are expressed by the Lamé constants λ and μ, which connect stress and strain in Hooke's law. For many rocks, these constants are almost equal; therefore we use $\lambda = \mu$ with μ being the rigidity. An elastic solid with this property is called a *Poisson solid*. Because the strain is dimensionless, μ has the same dimension as the stress. In the present study, we use $\mu = 30\,\text{GPa}$.
2. The (static) Green's function $G(\vec{y}_1, \vec{y}_2)$ defines the static response of the half space at a position $\vec{y}_1$ to a displacement at $\vec{y}_2$, which may arise from (coseismic) slip or (aseismic) creep motion. Due to the discretization of the fault plane into computational cells, we use the Green's function for static dislocations on rectangular fault patches of width dx and height dz, which is given in [20] and [47]. The main difference between this Green's function and the nearest-neighbor interaction of most slider-block models and cellular automata is the infinite-range interaction following a decay according to $1/r^3$, where r is the distance between source cell and receiver point.

Interseismic Processes

The motion of the tectonic plates, indicated in Fig. 2, is responsible for the build-up of stress in the fault zone. Geodetic measurements of surface displacements provide estimates of the velocity of the plates. For the San Andreas fault, a value of $v_{\text{pl}} = 35\,\text{mm/year}$ as a long-term average [55] is widely accepted and is adopted for the model. The displacement du in the regions surrounding the grid during a time period Δt is simply $\mathrm{d}u = v_{\text{pl}} \cdot \Delta t$. While the average slip rate $\dot{u}$ is independent of the location of a cell, the stress rate $\dot{\tau}$ depends on space. The assumption that the computational grid is embedded in a half-plane which undergoes constant creep, implies that cells at the boundaries of the grid experience higher load than cells in the center of the grid. The Green's function $G(i, j; k, l)$ defines the interaction of points (i, j) and (k, l) in the medium. In particular, the stress response at a position (i, j) to a static change of the displacement field d$u(k, l)$ is given by

$$\mathrm{d}\tau(i, j) = - \sum_{(k,l)\in \text{half space}} G(i, j;\, k, l) \cdot \mathrm{d}u(k, l)\,, \qquad (3)$$

where the minus sign stems from the fact that forward (right-lateral) slip of regions around a locked fault segment is equivalent to back (left-lateral) slip of the locked fault segment. Taking into account that

$$\sum_{(k,l)\in \text{half space}} G(i, j;\, k, l) = 0\,, \qquad (4)$$

Eq. (3) can be written as

$$\tau(i, j;\, t) = - \sum_{(k,l)\in \text{half space}} G(i, j;\, k, l)\cdot[u(k, l;\, t) - v_{\text{pl}} t]\,, \qquad (5)$$

where $u(k, l;\, t)$ is the total displacement at position (k, l) and time t since the start of the simulation. Because the surrounding regions sustain stable sliding, $u(k, l;\, t) = v_{\text{pl}} t$ for $(k, l) \notin$ grid, the slip deficit outside the fault region vanishes and it is sufficient to perform the summation on the computational grid:

$$\tau(i, j;\, t) = \sum_{(k,l)\in \text{grid}} G(i, j;\, k, l)\cdot[v_{\text{pl}} t - u(k, l;\, t)]\,. \qquad (6)$$

Equation (6) can be decomposed to a part for the tectonic loading and a residual part for slip on the computational grid. The tectonic loading follows the formula

$$\tau_{\text{load}}(i, j;\, t) = \gamma(i, j) \cdot t \qquad (7)$$

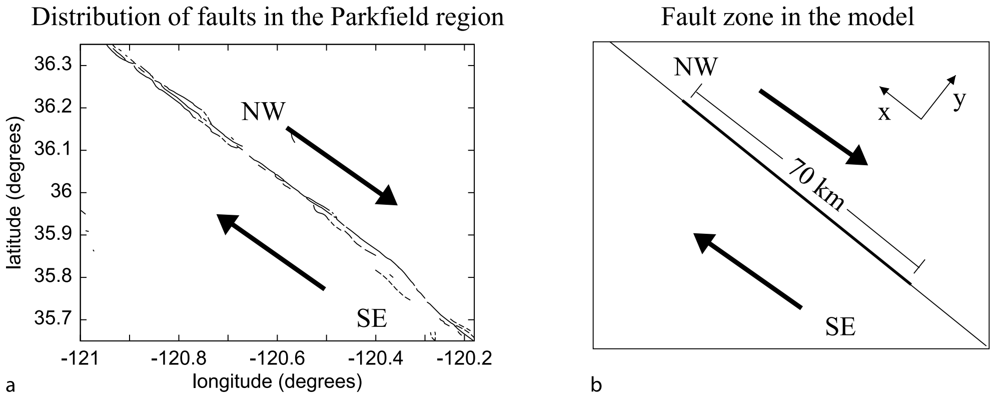
with a space-dependent but time-independent loading rate

$$\gamma(i, j) = v_{\text{pl}} \cdot \sum_{(k,l)\in \text{grid}} G(i, j;\, k, l)\,. \qquad (8)$$

The build-up of stress may be reduced by aseismic creep motion, which is implemented by a local constitutive law corresponding to lab-based dislocation creep [5]:

$$\dot{u}_{\text{creep}}(i, j;\, t) = c(i, j) \cdot \tau^3(i, j;\, t) \qquad (9)$$

with space dependent but time-independent creep coefficients $c(i, j)$.



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 2
a Distribution of faults in the Parkfield (California) region; b fault region in the model

Friction and Coseismic Stress Transfer; Quasidynamic Approach

It is widely accepted that earthquakes on large faults are due to frictional processes on pre-existing structures. The friction is therefore an important empirical ingredient of a fault model [56]. Numerous laboratory experiments have been carried out to characterize frictional behavior of different materials (see e. g. [19]). An important finding is that the friction coefficient defined as the ratio of shear stress τ_{shear} and compressional normal stress τ_{normal}, $\mu_{\text{f}} = \tau_{\text{shear}}/\tau_{\text{normal}}$ at the initiation of slip is approximately constant for many materials; the value of μ_{f} lies between 0.6 and 0.85. This observation, known as *Byerlee's law*, is related to the Coulomb failure criterion [16] for the Coulomb stress *CS*,

$$CS = \tau_{\text{shear}} - \mu_{\text{f}}\tau_{\text{normal}} \,. \tag{10}$$

The Coulomb stress depends on a plane where shear stress and normal stress are calculated. Neglecting cohesion, the Coulomb criterion for brittle failure is

$$CS \geq \tau_0 \,, \tag{11}$$

which for $CS = 0$ is Byerlee's law.

The North-American plate and the Pacific plate move in opposite directions along the fault plane having strike-slip motion. The absence of normal and thrust faulting reduces the problem to a one-dimensional motion: all parts of the fault move along the fault direction. The stress state of the fault is fully determined by the shear stress τ_{xy} in the coordinates given in Fig. 2b. Slip is initiated if τ_{xy} exceeds $\mu_{\text{f}}\tau_{yy}$. This quantity, which is called the static strength τ_{s} is constant in time if μ_{f} is assumed to be constant. Note that the normal stress on a planar fault in a homogeneous solid does not change [1]. The shear stress τ_{xy} will be denoted simply by τ. In this notation, the failure criterion Eq. (11) reduces to

$$\tau \geq \tau_{\text{s}} \,. \tag{12}$$

When a cell (k, l) fails, the stress drops in this cell to the arrest stress τ_{a}:

$$\tau(k, l) \rightarrow \tau_{\text{a}} \,, \tag{13}$$

where the value τ_{a} maybe space-dependent. The stress change produces a corresponding slip

$$\text{d}u(k, l) = \frac{\tau(k, l) - \tau_{\text{a}}}{G(k, l\,;\, k, l)} \tag{14}$$

with the self-stiffness $G(k, l\,;\, k, l)$ of cell (k, l).

The observational effect of dynamic weakening includes also a strength drop from the static strength to a lower dynamic strength:

$$\tau_{\text{s}} \rightarrow \tau_{\text{d}} \,. \tag{15}$$

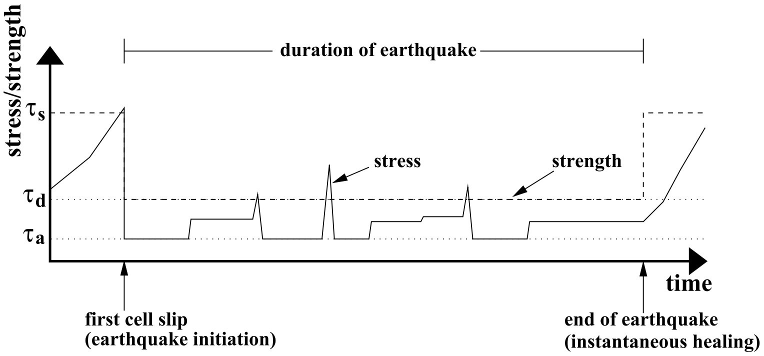
In particular, slipping material becomes weaker during rupture and recovers to the static level at the end of the rupture. This approximation of the strength evolution is known as static-kinetic friction.

The values τ_{s}, τ_{d}, and τ_{a} are connected by the dynamic overshoot coefficient D:

$$D = \frac{\tau_{\text{s}} - \tau_{\text{a}}}{\tau_{\text{s}} - \tau_{\text{d}}} \,, \tag{16}$$

or alternatively by the dynamic weakening coefficient ε:

$$\varepsilon = 1 - \frac{\tau_{\text{d}}}{\tau_{\text{s}}} \,. \tag{17}$$



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 3
Pictorial evolution of stress (*solid line*) and strength (*dashed line*) of a hypocenter cell in the quasidynamic approach

Following [10] we use in most simulations $D = 1.25$.

The redistribution of the stress release $\Delta\tau(k,l) = \tau(k,l) - \tau_a$ from cell (k,l) to a point (i,j) at time t is

$$\mathrm{d}\tau(i,j;t) = G(i,j;k,l) \cdot \delta\left(t - \frac{r(i,j;k,l)}{v_s}\right) \cdot \frac{\Delta\tau(k,l)}{G(k,l;k,l)}, \tag{18}$$

where $\delta(x)$ denotes the δ-function, which is 1 for $x = 0$ and 0 else; v_s is the shear-wave velocity, and $r(i,j;k,l)$ is the distance between source cell (k,l) and receiver position (i,j). That is, regions far from the slipping cell receive their stress portion later than regions close to the slipping cell. The value of v_s is assumed to be constant. Each "stress transfer event" associated with Eq. (18) gives a transfer of a stress $\mathrm{d}\tau$ from a source cell (k,l) to a receiver cell (i,j) at time t. This time-dependent stress transfer is called the *quasidynamic* approach in contrast to the *quasistatic* approach used in most similar models.

The evolution of stress and strength in a failing cell is shown schematically in Fig. 3. When the slip is initiated, both the stress and the strength drop. Due to coseismic stress transfer during the event, the cell may slip several times before the earthquake is terminated and instantaneous healing takes place in all cells. The piecewise constant failure envelope (dashed line) indicates static-kinetic friction. A model version with gradual healing was employed by [79]. A review of analytical results associated with the basic model in the context of a large universality class is given in ▶ Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of.

We note that the Green's function leads to an infinite interaction range. Using open boundary conditions with respect to the computational grid, the stress release from a slipping cell is not conserved on the grid, but on the infinite half plane.

Data

The model produces two types of data, earthquake catalogs and histories of stress and displacement fields. As demonstrated below, all parameters of the model have physical dimensions and can therefore be compared directly with real data, where they are available. This is in contrast to most of the slider-block and cellular automaton models.

Earthquake catalogs include values of the earthquake time, hypocenter, and size. The time of an earthquake is the time of the first slip; the hypocenter is determined by the position of the corresponding cell along strike and depth. The size of an event can be described by different measures: The rupture area A is the total area, which slipped during an earthquake. The potency

$$P = \int u(A)\mathrm{d}A \tag{19}$$

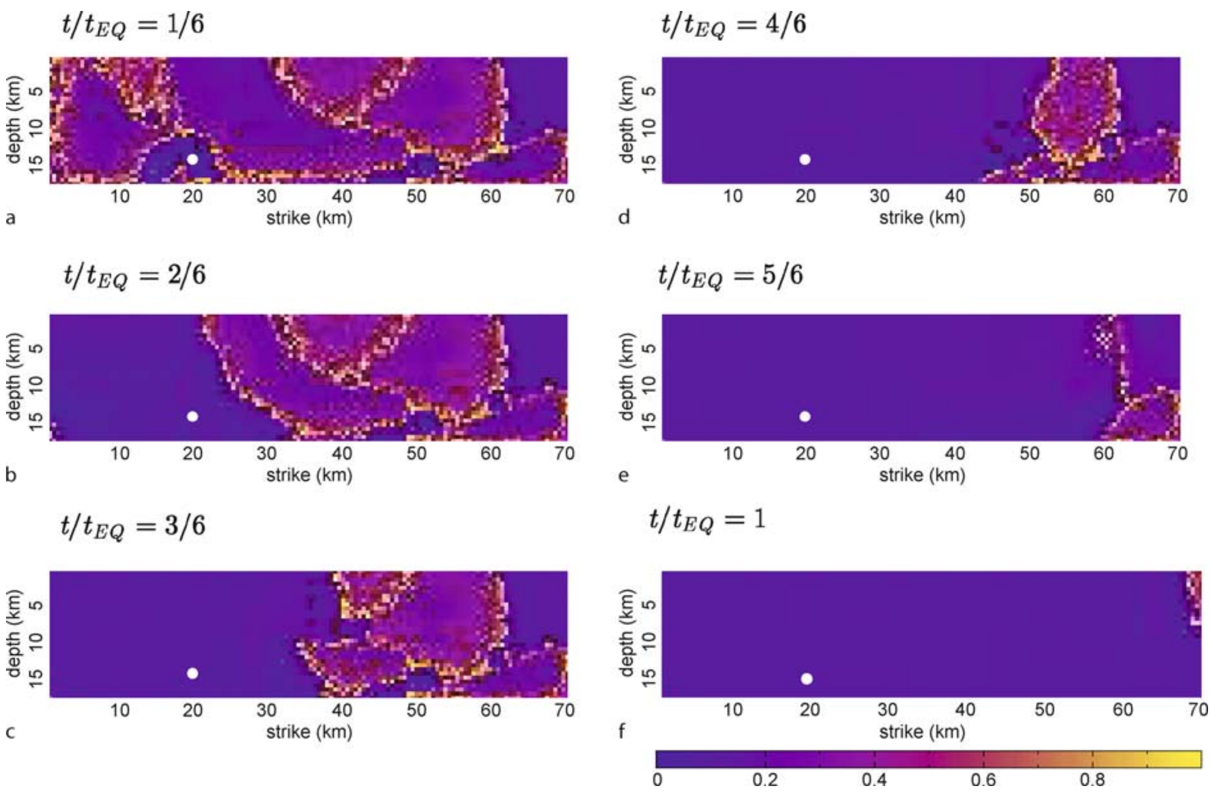
measures [7] the total slip during the event and is related to the seismic moment m_0 by the rigidity: $m_0 = \mu P$. The (moment) magnitude M can be calculated from the potency [10] using

$$M = (2/3)\log_{10}(P) + 3.6\,, \tag{20}$$

where P is given in $\mathrm{cm} \cdot \mathrm{km}^2$.

Results

Numerous simulations of the model described in the previous section have been performed. Firstly, simulations have been examined with respect to the spatiotemporal propagation of stress during single earthquakes ("rupture histories"). Then, long deformation histories have been



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 4

Snapshots of rupture evolution for a system-wide event on a smooth fault without creep motion. *t* denotes the time after the rupture initiation (given in units of the total earthquake duration t_{EQ}). The white circle is the hypocenter of the event. The figure shows the dimensionless stress state $\hat{\tau} = \frac{\tau - \tau_a}{\tau_s - \tau_a}$ of the cells. a $t/t_{\mathrm{EQ}} = 1/6$; b $t/t_{\mathrm{EQ}} = 2/6$; c $t/t_{\mathrm{EQ}} = 3/6$; d $t/t_{\mathrm{EQ}} = 4/6$; e $t/t_{\mathrm{EQ}} = 5/6$; f $t/t_{\mathrm{EQ}} = 1$

simulated in order to search in a large fraction of the parameter space for relationships between input parameters and observed seismicity features. In this section, a selection of key results is presented and discussed in relation to critical states of seismicity.

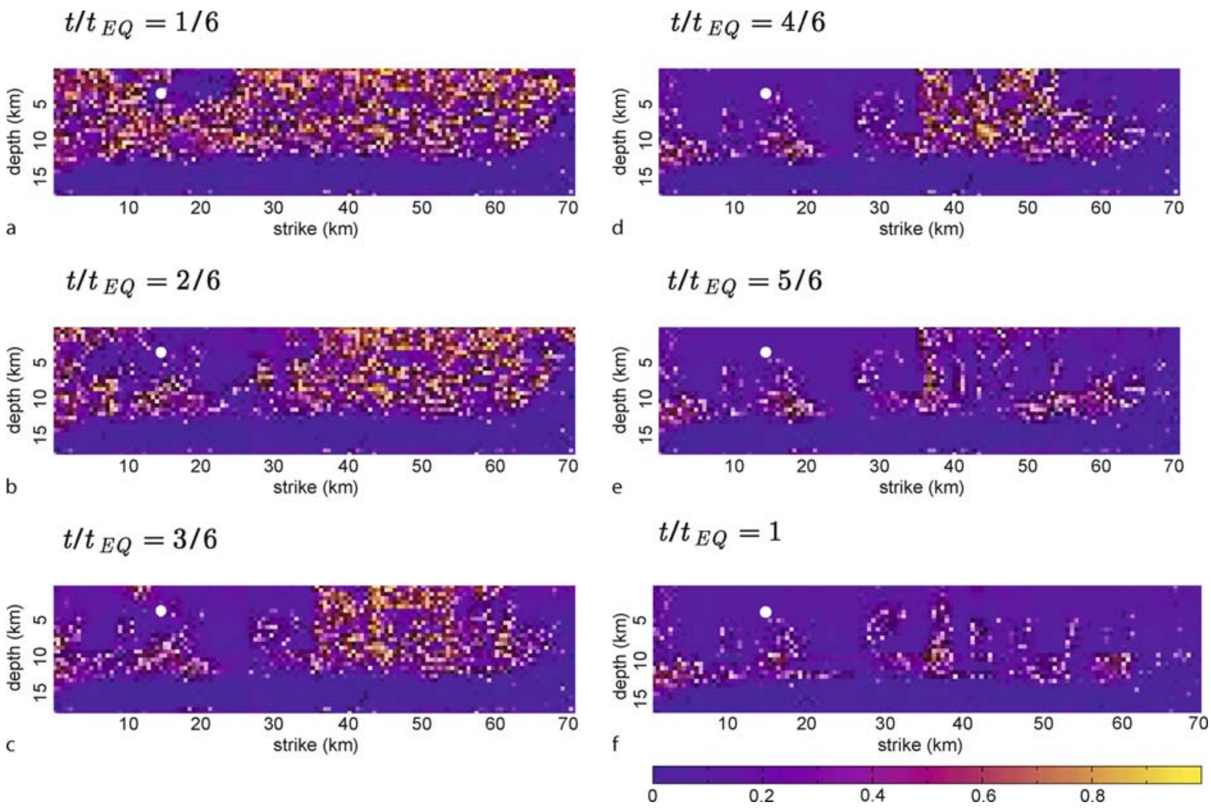
Rupture Histories

We compare qualitatively rupture histories of large earthquakes for three end-member cases in parameter space:

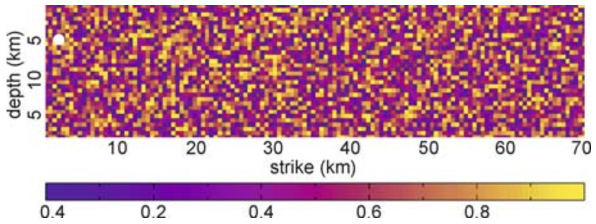
(1) a smooth fault,
(2) a rough fault, and
(3) a fault without dynamic weakening ($\tau_d = \tau_s$ or $D \to \infty$ in Eq. (16)).

Following [5], we vary the degree of quenched spatial disorder for a particular realization by introducing barriers of high stress drop $\tau_s - \tau_a$ in an environment of low stress drop.

The observation that smooth faults show a more regular earthquake occurrence than rough faults, can be explained by the ability of the stress field to synchronize on certain fault patches. On a disordered fault, this type of synchronization is unlikely. Figure 4a shows the stress field (normalized between 0 and 1) immediately before a large earthquake on a smooth fault. The most striking feature is the emergence of clearly defined patches with highly loaded boundaries. During rupture evolution, these patches rupture almost in series until the fault is nearly unloaded (see Fig. 4b–f). A different situation is shown in Fig. 5, corresponding to a rough fault with creep coefficients $c(i, j)$ (Eq. (9)) that increase with depth leading to a brittle-ductile transition zone as in [5] and [80]. Here the stress field in the brittle regime is irregular without obvious pattern formation. Similar behavior is found in a case where dynamic weakening is switched off ($D \to \infty$); in other words, the material heals instantaneously. Figure 6 shows the stress field in this case. In analytical studies, it has been shown that this corresponds exactly to a critical point in a phase diagram spanned by the stress dissipation and dynamic weakening [22,27]. Observational results indicate [68] that irregular slip histories and power



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 5
Same as Fig. 4 for a strongly disordered fault with a brittle-ductile transition at about 15 km depth. a $t/t_{\text{EQ}} = 1/6$; b $t/t_{\text{EQ}} = 2/6$; c $t/t_{\text{EQ}} = 3/6$; d $t/t_{\text{EQ}} = 4/6$; e $t/t_{\text{EQ}} = 5/6$; f $t/t_{\text{EQ}} = 1$



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 6
Same as Fig. 4a for a fault without dynamic weakening ($\tau_d = \tau_s$) corresponding to a dynamic overshoot coefficient $D \to \infty$ (Eq. (16))

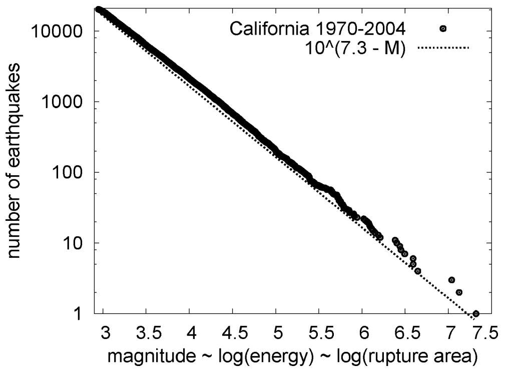
law frequency-size distributions are associated with geometrically disordered fault structures, while characteristic earthquake statistics and overall regular ruptures are found on mature fault with large total displacements.

Although the stress field shows a complex evolution during a simulation, the presence or absence of characteristic length scales indicating the relation to a critical point is easily detected. From an observational point of view, the stress field is not accessible. The evolution of the displacement field may be estimated, e. g. from seismic and geodetic data using slip inversion techniques. Because of the high uncertainties in the calculated slip histories, a quantitative comparison of the simulated data with "natural" data is questionable. However, general features of the quasidynamic ruptures are quite realistic, e. g. the irregular patterns in Fig. 5 resemble the rupture of the Chi–Chi (Taiwan) earthquake on September 21, 1999 ($M_w = 7.6$) [58].

Later we will show that the frequency-size distribution of earthquakes can serve to some extent as a proxy for the degree of disorder of the stress field. Ben-Zion et al. [12] discuss additional seismicity functions that may be used as surrogate variables for the stress.

Frequency–Size Distributions

The frequency-size (FS) distribution is one of the most important characteristics of observed seismicity. For world-



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 7

Frequency-size distribution for California from 1970 to 2004; the dashed line denotes a power law fit to the data

wide seismicity, this distribution is given by the Gutenberg–Richter law (Eq. (1)). Figure 7 shows the FS distribution of California seismicity from 1970 to 2004. Here we use the non-cumulative version of Eq. (1), where N is the number of earthquakes with magnitude between M and $M + \mathrm{d}M$ with a magnitude bin $\mathrm{d}M = 0.1$.

For individual fault zones the FS distribution can deviate from Eq. (1), especially for high magnitudes. Examples are given in Fig. 8, which shows the FS distribution of the Parkfield segment (Fig. 8a) and for the San Jacinto fault (Fig. 8b) in California for a time span of 45 years. The distribution of the Parkfield segment consists of two parts: A scaling regime for $2.2 \leq M \leq 4.5$ and a "bump" for $4.5 < M \leq 6.0$. For the San Jacinto fault, the scaling range is observed for almost all events ($2.2 \leq M \leq 5.0$). The decrease for $M \approx 2$ in both plots is probably due to catalog incompleteness.

The FS distribution as shown in Fig. 8a is called the characteristic earthquake distribution, because of the increased probability for the occurrence of large ("characteristic") events compared to the prediction of the Gutenberg–Richter relation. The latter is an exponential distribution for the earthquake frequency as a function of magnitude, or a power law distribution for the earthquake frequency as a function of potency (Eqs. (19), (20)), moment, energy, or rupture area, over a broad range of magnitudes [66]. The Gutenberg–Richter relation is "scale-free" because a power law distribution indicates the absence of a characteristic scale of the earthquake size [64]. In terms of critical point processes, the absence of a characteristic length scale indicates that the system is close to the critical point. In this state, earthquakes of all magnitudes can occur, or each small rupture can grow into a large one. Therefore, the frequency-size distribution can serve as a proxy for the current state of a system in relation to a critical point.

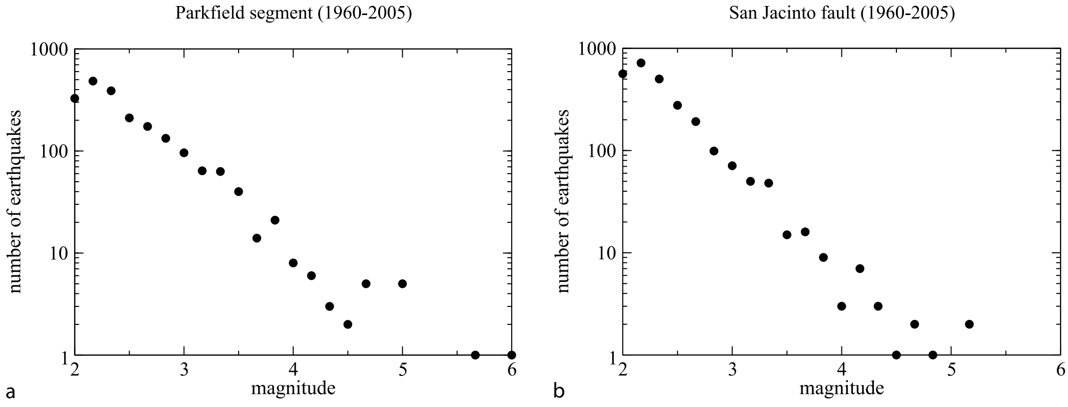
The FS distribution in a model can be tuned by varying the mean stress $\langle\tau\rangle$ on the fault, where $\langle\rangle$ denotes the spatial average. This can be achieved, for instance, by varying brittle properties, e. g. in terms of the dynamic overshoot coefficient D (Eq. (16)), or by introducing dissipation [31,79]. Figure 9 shows FS distributions for two different values of $D : D = 5/4$ (Fig. 9a) and a higher value $D = 5/3$ (Fig. 9b). While Fig. 9a follows a characteristic earthquake behavior similar to the Parkfield case (Fig. 8a), Fig. 9b resembles the shape of the FS distribution of the San Jacinto fault (Fig. 8b).

As an outcome, three cases can be distinguished by means of a critical mean stress τ_{crit}:

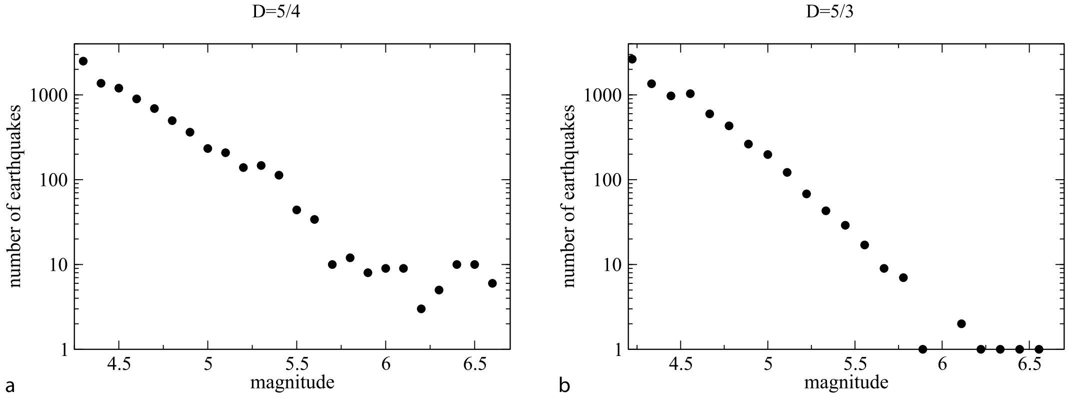
1. subcritical fault ($\langle\tau\rangle < \tau_{\text{crit}}$): the mean stress on the fault is too small to produce large events. The system is always far from the critical point. The FS distribution is a truncated Gutenberg–Richter law.
2. supercritical fault ($\langle\tau\rangle > \tau_{\text{crit}}$): the mean stress is high and produces frequently large events. After a large earthquake (critical point), the stress level is low (system is far from the critical point) and recovers slowly (approaches the critical point). The FS distribution is a characteristic earthquake distribution.
3. critical fault ($\langle\tau\rangle \approx \tau_{\text{crit}}$): the system is always close to the critical point with scale-free characteristics. The FS distribution is a Gutenberg–Richter law with a scaling range over all magnitudes.

If the FS distribution is plotted as a function of the model parameters, the result can be visualized by a phase diagram [22,31,79,80]. An example is provided in Fig. 10, which shows schematically the phase diagram spanned by the degree of quenched spatial disorder and $1 - \varepsilon$ with the dynamic weakening coefficient ε (Eq. (17)). The phase diagram summarizes results from various studies, which demonstrate that the degree of spatial disorder of the stress drop acts as a tuning parameter for the FS distribution [5,36,80].

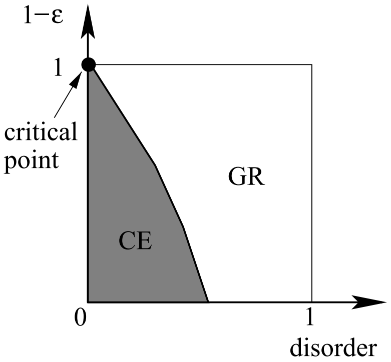
If the model is in the transition regime between Gutenberg–Richter statistics and characteristic earthquake behavior, the ability of the stress field to synchronize on parts of the fault can have additional impact on the dynamics of seismicity: for a model with small cells and high stress fluctuations along the cell boundaries arising from a high degree of spatial disorder, the system can undergo a spontaneous transition from an ordered state and characteristic behavior to a disordered state following Gutenberg–Richter statistics. Due to the high fluctuations in the



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 8
Frequency-size distribution for two faults in California: a the Parkfield segment, and b the San Jacinto fault

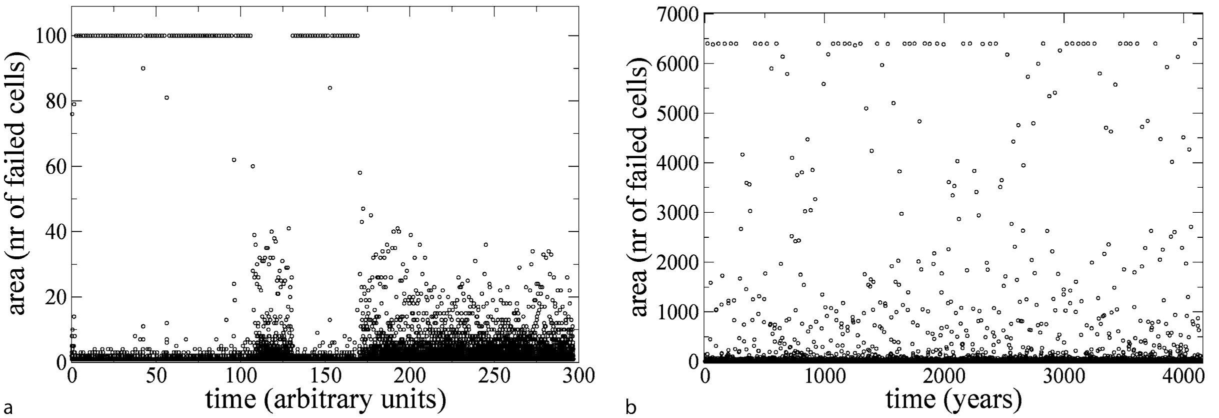


Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 9
Frequency-size distribution for model realizations with different dynamic overshoot coefficients (Eq. (16)): a $D = 5/4$, b $D = 5/3$

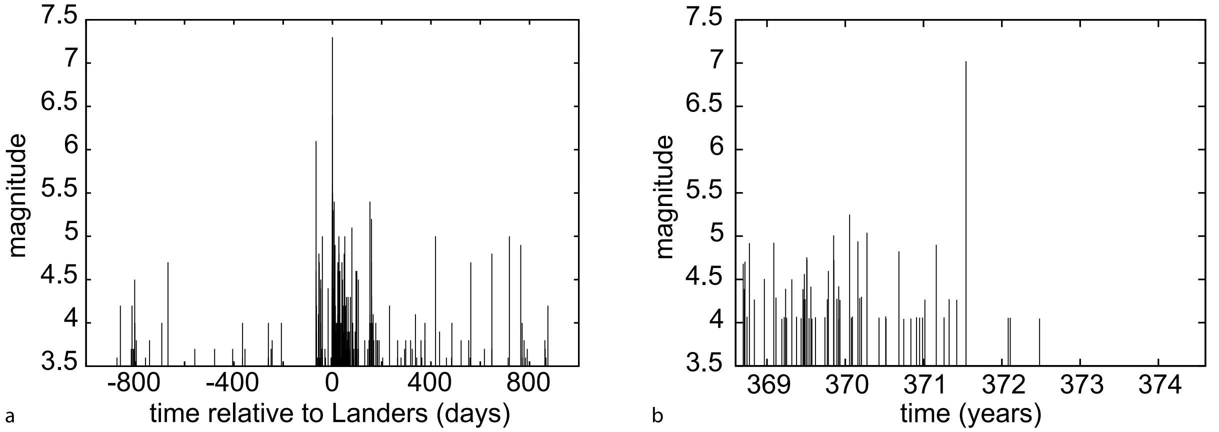


Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 10
Phase diagram for the frequency-size distribution (GR = Gutenberg–Richter distribution, CE=characteristic earthquake distribution) spanned by the degree of quenched spatial disorder and the dynamic weakening represented by ε. The upper left corner corresponds exactly to a critical point [22,27] and results in scale-free characteristics as shown in Fig. 6

stress field, there is some probability that a certain number of cells synchronize by chance, leading to an ordered behavior for some seismic cycles, until the order is destroyed, again resulting from stress fluctuations. This type of mode-switching behavior has been observed earlier in a mean-field model and a damage rheology model [11,22]. Figure 11a gives a corresponding earthquake sequence with spontaneous mode-switching behavior. Figure 11b shows a sequence calculated with a higher grid resolution (128 × 50 cells). The tendency to mode-switching is less pronounced, but still visible. In [79] it is shown that the emergence of such mode-switching depends both on the spatial range of interaction (given as the decay of the Green's function) and the discretization of the computational grid. In the less realistic model of [22], where the stress redistribution is governed by a constant (space-independent) Green's function, analytical expressions for persistence times have been calculated [27]. In [11] some evidence for mode-switching behavior in long seismic



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 11
Earthquake area (measured as the number of failed cells) as a function of time a in the mean-field model of Dahmen et al. [22] for a fault with 100 cells and b in the elastic model with 128 × 50 cells



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 12
Earthquakes before and after a mainshock: a the *M*7.3 Landers (California) earthquake; b *M*7.3 earthquake in the basic version of the model

records based on paleoseismic and geologic data from the Dead Sea fault and other regions are discussed. However, the relevance of mode-switching to natural seismicity remains unclear due to the general lack of very long data records.

Aftershocks and Foreshocks

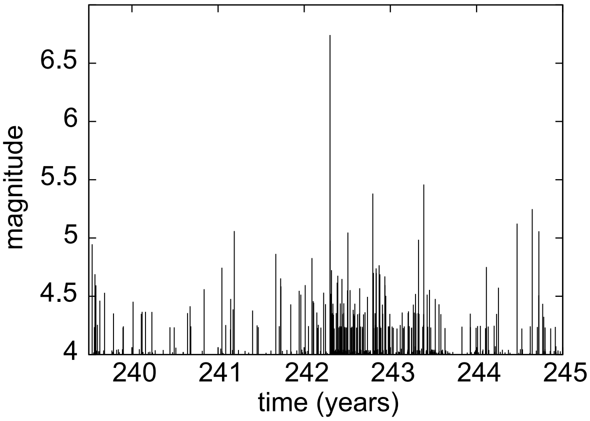
The most pronounced temporal pattern in observed seismicity is the emergence of strongly clustered aftershock activity following a large earthquake. Apart from the Omori–Utsu law (Eq. (2)), it is widely accepted that aftershocks are characterized by the following properties:

1. The aftershock rate scales with the mainshock size [51].
2. Aftershocks occur predominantly around the edges of the ruptured fault segments [66].
3. Båth's law [4]: The magnitude of the largest aftershock is usually about one unit smaller than the mainshock magnitude.

Deviations from the Omori–Utsu law, especially for rough faults, are discussed in [45]. While aftershocks are observed after almost all large earthquakes, foreshocks occur less frequent [71]. As a consequence, much less is known about the properties of these events. Kagan and Knopoff [40] and Jones and Molnar [39] propose a power law increase of activity according to an "inverse" Omori–Utsu law.

Figure 12a shows an example for the aftershock sequence following the *M*7.3 Landers earthquake in California on June 28, 1992. An earthquake of similar size generated by the model is given in Fig. 12b. The absence of aftershocks in the simulation is clearly visible. The reason for the lack of aftershocks is the unloading of the fault resulting from the mainshock: When a large fraction of the fault has ruptured, the stress in this region will be close to the arrest stress after the event. Consequently, the seismicity rate will be almost zero until the stress field has recovered to a moderate level.

Discussions for likely mechanisms of aftershocks are given in [9] and [81]. A common feature is the presence of rapid postseismic stress which generates aftershock activity. In [32], for instance, postseismic stress has been attributed to a viscoelastic relaxation process following the mainshock. In the work discussed here, continuous creep displacement following Eq. (9) is assumed. Additionally, the computational grid is divided by aseismic barriers from the free surface to depth into several seismically active fault segments. As shown in [81], this modification results in a concentration of stress in the aseismic regions during rupture and, subsequently in a release of stress after the event according to the coupled creep process. This stress release triggers aftershock sequences obeying the Omori–Utsu law (Eq. (2)). A typical aftershock sequence after a $M6.8$ event is shown in Fig. 13. In agreement with Båth's law, the strongest aftershock has the magnitude $M = 5.5$ in this sequence. The sequence shows also the effect of secondary aftershocks, namely aftershocks of aftershocks [61]. The stacked earthquake rate as a function of the time after the mainshock is given in Fig. 14. In this case, where the barriers are characterized by creep coefficients higher by a factor of 10^5 than in the other patches, a realistic Omori exponent of $p = 1$ is found.



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 13
Earthquakes before and after a mainshock with $M = 6.8$ in the modified model

Aftershock sequences like in Fig. 13 emerge after all large events in the extended model. In contrast, there is generally no clear foreshock signal visible in single sequences. However, stacking many sequences together unveils a slight increase of the earthquakes rate prior to a mainshock supporting the observation of accelerating foreshock activity. An explanation of these events can be given in the following way: Between two mainshocks the stress field organizes itself towards a critical state, where the next large earthquake can occur. This critical state is characterized by a disordered stress field and the absence of a typical length scale, where earthquakes of all sizes can occur [12]. The mainshock may occur immediately or after some small to moderate events. The latter case can be considered as a single earthquake, which is interrupted in the beginning. This phenomenon of delayed rupture propagation has also provided a successful explanation of foreshocks and aftershocks in a cellular automaton model [33,35].

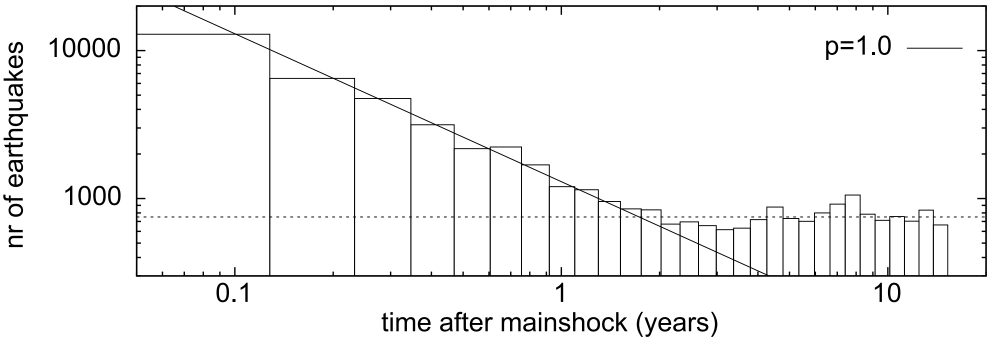
The hypothesis that foreshocks occur in the critical point and belong, in principle, to the mainshock, can be verified by means of the findings from Subsect. "Frequency–Size Distributions". In particular, the frequency-size distribution in the critical point (or close to the critical point) is expected to show scale-free statistics. If an overall smooth fault model following characteristic earthquake statistics is studied over a long time period, the approach of the critical point should be seen in terms of a change of the frequency-size distribution towards Gutenberg–Richter behavior [12]. This change of frequency-size statistics is indeed observed in the model (Fig. 15) and supports the validity of the critical point concept [82].

Accelerating Moment Release

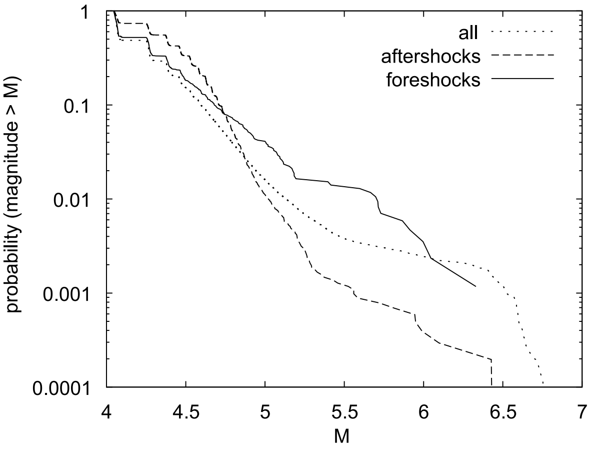
In the previous section, it has been argued that large earthquakes are associated with a critical point and the preparation process is characterized by increasing disorder of the stress field and increasing tendency for scale-free characteristics in the frequency-size distribution. Further support for critical point dynamics has been provided by the observational finding of [17] that the cumulative Benioff strain $\Sigma\Omega(t)$ follows a power law time-to-failure relation prior to the $M7.1$ Loma Prieta earthquake on October 17, 1989:

$$\Sigma\Omega(t) = \sum_{i=1}^{N(t)} \sqrt{E_i} = A - B(t_\mathrm{f} - t)^m \qquad (21)$$

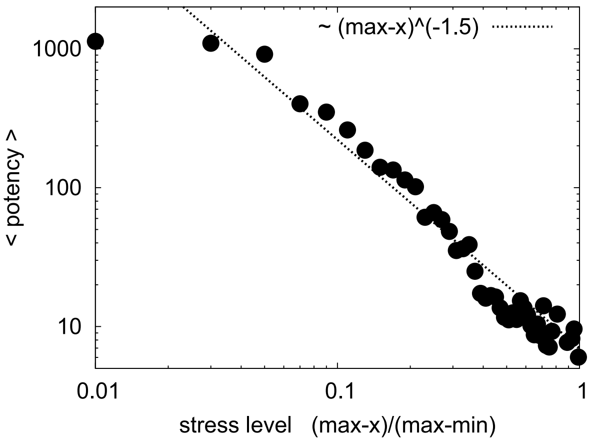
Here, E_i is the energy release of earthquake i, $N(t)$ is the number of earthquakes before time t; t_f is the failure time



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 14
Earthquake rate as a function of time for the model with seismic and aseismic regions. The calculation is based on a simulation with 200,000 earthquakes covering about 5000 years; the earthquake rates are averaged over about 300 mainshocks. A fit of the Omori–Utsu law (Eq. (2)) with $p = 1$ is denoted as a solid line. The dashed line gives the estimated background level of seismicity



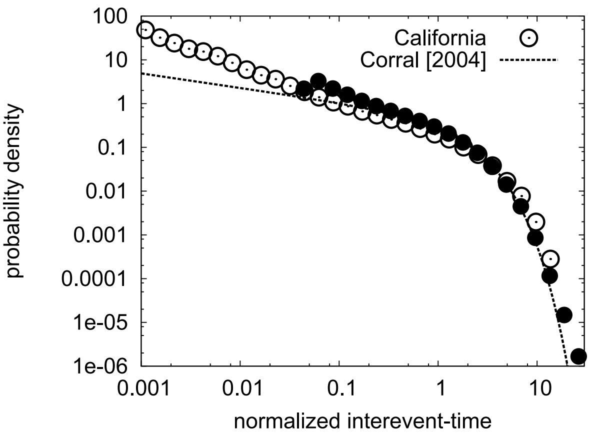
Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 15
Frequency-magnitude distributions of all earthquakes, foreshocks and aftershocks. Foreshocks and aftershocks are defined as earthquakes occurring within one month before and after an earthquake with $M \geq 6$



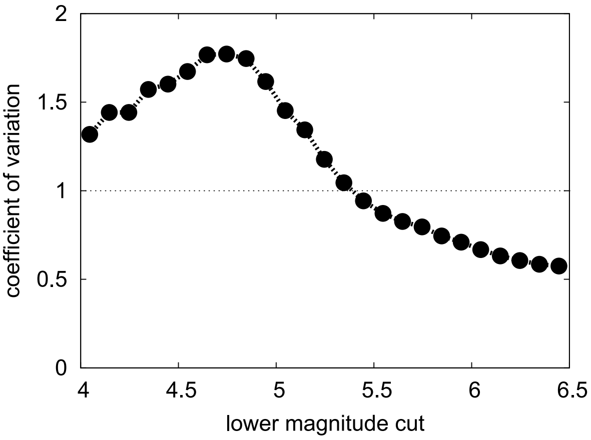
Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 16
Mean potency release (Eq. (19)) as a function of the stress level. The stress level is normalized to the maximum (max) and minimum (min) observed stress

and A, B, and $m > 0$ are constants. Similar studies for numerous seismically active regions followed (see [8,77] and references therein).

The time-to-failure relation Eq. (21) has been explained by [54] and [60] from the viewpoint of renormalization theory and by [8] and [65] in terms of damage rheology. Similar to the findings about foreshocks, the time-to-failure pattern is not universal. Therefore, a stacking procedure is adopted in order to obtain robust results on the validity of Eq. (21) in the model. This is not straightforward, since the interval of accelerating moment release is not known a priori and the duration of a whole seismic cycle, as an upper limit, is not constant. To normalize the time interval for the stacking, the potency release (Eq. (19)) is computed as a function of the (normalized) stress level (Fig. 16). Taking into account that the stress level increases almost linearly during a large fraction of the seismic cycle, the stress level axis in Fig. 16 can effectively be replaced by the time axis leading to a power law dependence of the potency release on time. The best fit is provided with an exponent $s = -1.5$. Transforming the potency release to the cumulative Benioff strain (Eq. (21)), results in an exponent $m = 0.25$ in Eq. (21). This finding is based on a simulation over 5000 years; the exponent is in good agreement with the theoretical work [53], that derives $m = 0.25$ for a spinodal model, and the analytical result of $m = 0.3$ in the damage mechanics model [8]. An observational study of California seismicity finds m between 0.1 and 0.55 [15].



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 17
The normalized interevent time distribution of the model simulations (black dots) compared with the result of [21] and the distribution of earthquakes in California (ANSS catalog of $M \geq 3$ earthquakes occurred between 1970 and 2004 within 29° and 43° latitude and −113° and −123° longitude)



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 18
The temporal earthquake occurrence quantified by the coefficient of variation as a function of the lower magnitude cutoff. Values larger than 1 indicate clustering, whereas lower values point to quasiperiodic behavior

Interevent Times

In recent studies it has been shown that the distribution of interevent times can be described by a universal law. In particular, the distributions from different tectonic environments, different spatial scales (from worldwide to local seismicity) and different magnitude ranges collapse if the time Δt is rescaled with the rate R_{xy} of seismic occurrence in a region denoted by (x, y) [21]. Such rescaling leads to

$$D_{xy}(\Delta t) = R_{xy} \cdot f(R_{xy}\Delta t) , \tag{22}$$

where D_{xy} is the probability density for the interevent time Δt, and f can be expressed by a generalized gamma distribution

$$f(\theta) = C\frac{1}{\theta^{\gamma-1}} \exp\left(-\frac{\theta^{\delta}}{B}\right) . \tag{23}$$

The parameters C, γ, δ, and B have been determined by a fit to several observational catalogs [21].

In Fig. 17 we compare $D_{xy}(\Delta t)$ from Eq. (22) with two earthquake catalogs: (1) The ANSS catalog of California (catalog ranges are given in the caption), and (2) the model catalog. Due to the universality of Eq. (22) with respect to different spatial scales, the comparison of the model simulating a single fault of 70 km length with a region of hundreds of kilometers including several faults in California does not require coarse graining the ANSS catalog. In the region where the interevent times are calculated, we find a remarkable agreement of the three curves. For small values of Δt, Eq. (22) deviates from the California data; for high values the model has a slightly better correspondence with the observational data than Eq. (22). Thus the results generally support the recent findings of [21].

The degree of temporal clustering of earthquakes can be estimated by the coefficient of variation CV of the interevent time distribution.

$$CV = \sigma/\mu , \tag{24}$$

where σ is the standard deviation and μ the mean value of the interevent time distribution. Values of $CV > 1$ denote clustered activity, while $CV < 1$ represents quasiperiodic occurrence of events. The case $CV = 1$ corresponds to a random Poisson process. The studies of [5] and [80] have found that the clustering properties of the large events depend on the degree of quenched spatial disorder of the fault. Figure 18 shows that CV as a function of the lower magnitude cutoff has a characteristic shape. The values of CV are higher than 1 (clustered) for small and intermediate earthquakes ($M \leq 5.4$) and smaller than 1 (quasiperiodic) for larger earthquakes. This corresponds to the case of a low degree of disorder in [80], because the brittle cells which participate in an earthquake have no significant spatial disorder. We note that this behavior resembles the seismicity on the Parkfield segment of the San Andreas fault, which is characterized by a quasiperiodic occurrence of mainshocks. Based on the analysis of 37 earthquake sequences, an estimation of $CV \approx 0.5$ has been found for multiple tectonic environments [26].

A different behavior is observed on the San Jacinto fault in California, where the largest events occur less regularly and have overall smaller magnitudes. As discussed in [5] and [80], this can be modeled by imposing higher degrees of disorder leading to a broader range of spatial size scales, e. g. by using a higher number of near-vertical barriers. While barriers provide a simple and physically motivated way to tune the degree of disorder, other types of heterogeneities may work as well, as long as they are able to produce strong enough fluctuations of the stress field. As an example, we mention fractal distributions of the stress drop, which can be tuned easily by changing the fractal dimension [63,80].

Recurrence Times of Large Earthquakes

While interevent times are waiting times between successive earthquakes in a given catalog, recurrence times are defined as waiting times between two successive *large* events, typically in the magnitude range $6 \leq M \leq 9$, depending on the region. For example, on the Parkfield segment of the San Andreas fault seven $\sim M6$ earthquakes occurred between 1857 and 2004 with recurrence times $T_1 = 24$ years, $T_2 = 20$ years, $T_3 = 21$ years, $T_4 = 12$ years, $T_5 = 32$ years, and $T_6 = 38$ years.

The distribution of recurrence times of large earthquakes is crucial for the calculation of seismic hazard. Due to a lack of observational data, this distribution is unknown for real fault systems. Commonly used distributions are based on extreme value statistics and on models for catastrophic failure. These include the lognormal distribution [50], the Brownian passage time distribution [44], and the Gumbel distribution [29]. All distributions are characterized by a maximum for a certain recurrence time followed by an asymptotic decay. The Brownian passage time distribution and the lognormal distribution have been used by the Working Group on California Earthquake Probabilities [70], e. g. for calculating earthquake probabilities in the San Francisco Bay area.

Figure 19a shows the probability density function (pdf) of the recurrence times of earthquakes with magnitude $M > 6.2$ in a realization of the numerical model for the Parkfield region [74]. Since we focus on long time-scales, we use here a minimal model without aseismic creep and strong spatial heterogeneities. This model leads to characteristic earthquake statistics and quasiperiodic occurrence of large events, and can therefore serve as a model framework for large earthquakes on the Parkfield segment. However, quantities which are only poorly known from empirical data, e. g. the stress drop, have to be chosen in order to perform a numerical simulation. Starting with an imposed uniform a priori distribution $P(\Delta\tau)$ of stress drops between a lower bound $\Delta\tau_{\min}$ and an upper bound $\Delta\tau_{\max}$, an a posteriori distribution $P(\Delta\tau|T_1,\ldots,T_N)$ can be estimated using observational recurrence times $T_1,\ldots,T_N$ from Parkfield and Bayes' theorem [13],

$$P(\Delta\tau|T_1,\ldots,T_N) = \frac{P(T_1,\ldots,T_N|\Delta\tau)P(\Delta\tau)}{\sum\limits_{s=\Delta\tau_{\min}}^{\Delta\tau_{\max}} P(T_1,\ldots,T_N|s)P(s)}\,, \tag{25}$$

with the likelihood function

$$P(T_1,\ldots,T_N|\Delta\tau) = \prod_{i=1}^{N} f(T_i|\Delta\tau)\,. \tag{26}$$

The function $f(T_i|\Delta\tau)$ denotes the pdf of recurrence times simulated with a model stress drop $\Delta\tau$. To get an analytic expression of this function, it is fitted by a Gamma distribution $f(t) = \beta^{-1}(\Gamma(\gamma))^{-1}(\frac{t-\mu}{\beta})^{\gamma-1}\exp(-\frac{t-\mu}{\beta})$ with the location parameter μ, the shape parameter $\gamma \equiv 2.0$ and the scale parameter β (with $x \geq \mu\,;\ \gamma,\beta > 0$). For an example see Fig. 19a. In [74] it is shown that the mean value μ_t and the standard deviation σ_t of the fits in this model are related to the average stress drop of a large earthquake $\Delta\tau$ by the simple empirical relations

$$\begin{aligned} \mu_t(\Delta\tau) &= 9.7\cdot\Delta\tau \\ \sigma_t(\Delta\tau) &= 1.8\cdot\Delta\tau^2 - 6.8\cdot\Delta\tau + 11.7 \end{aligned} \tag{27}$$

with $\mu_t\,,\sigma_t$ in years and $\Delta\tau$ in MPa. Using this approximation in combination with six observational recurrence times from $\sim M6$ earthquakes on the Parkfield segment, we find the a posteriori distribution of stress drops shown in Fig. 19b. The position where this distribution reaches the maximum, $\Delta\tau = (3.04 \pm 0.27)$ MPa, is the most representative value of the stress drop of $\sim M6$ Parkfield events.

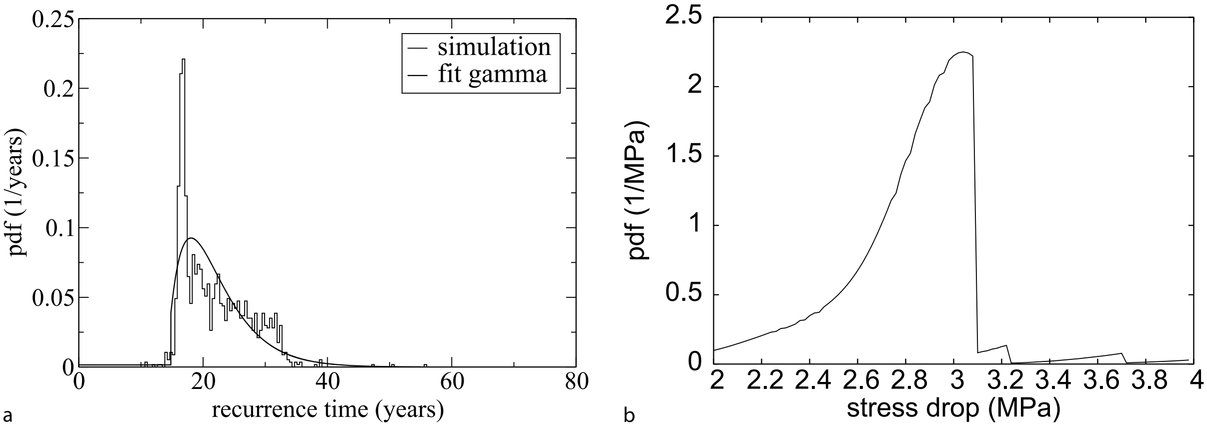
The cumulative probability density function (cdf) of recurrence times based on Eq. (26) and the observational data can now be calculated by

$$C(t) = \int_0^t \int_{\Delta\tau_{\min}}^{\Delta\tau_{\max}} f(t'|\Delta\tau)P(\Delta\tau|T_1,\ldots,T_N)\mathrm{d}\Delta\tau\,\mathrm{d}t'\,. \tag{28}$$

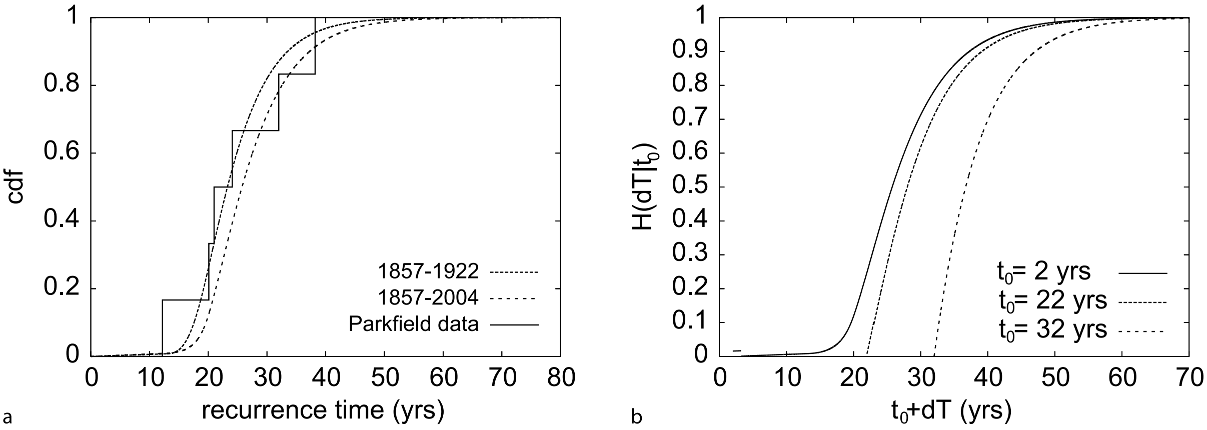
The hazard function

$$H(\Delta t|t_0) = \frac{C(t_0+\Delta t) - C(t_0)}{1 - C(t_0)} \tag{29}$$

is the conditional probability that the next large earthquake occurs in the interval $[t_0\,;\ t_0+\Delta t]$ given the time t_0 since the last large event. Results for two choices of observational data (corresponding to two different observa-



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 19
a Approximated probability density function of the recurrence time distribution of large earthquakes ($M > 6.2$) for a simulated earthquake catalog and fit with a truncated Gamma distribution; **b** A posteriori distribution $P(\Delta\tau|T_1, \ldots, T_N)$ of stress drop $\Delta\tau$ calculated with Bayes' theorem (Eq. (25))



Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 20
a Cumulative recurrence time distribution $C(t)$ (Eq. (28)) for (1) the Bayesian approach with three data points: long-dashed line; (2) the Bayesian approach with six data points: short-dashed line. The solid line denotes the cdf of the six Parkfield recurrence times; **b** Hazard function $H(t_0|\Delta T)$ (Eq. (29)) based on the six observational recurrence times between 1857 and 2004 **a** as a function of dT for different values of t_0

tional periods) in comparison to the Parkfield cdf are given in Fig. 20a. The hazard function for three fixed values of t_0 and varying Δt is given in Fig 20b.

This approach enables us to calculate the most likely occurrence time of the next (post 2004) Parkfield earthquake by picking the maximum of the (non-cumulative) recurrence time distribution (inner integral in Eq. (28)) after taking all Parkfield earthquakes (1857–2004) into account. Based on the analysis done so far, we may forecast the next $\sim M6$ Parkfield earthquake to occur in May 2027. The error associated with one standard deviation of the pdf is 7.7 years. We note, however, that the probability for the occurrence of a $\sim M6$ earthquake between May 2026 and May 2028 is only about 14%.

Summary and Conclusions

The present review deals with the analysis, the understanding and the interpretation of seismicity patterns with a special focus on the critical point concept for large earthquakes. Both physical modeling and data analysis are discussed. This study aims at practical applications to model data from real fault zones. A point of particular interest is the detection of phenomena prior to large earthquakes and their relevance for a possible prediction of these events. Despite numerous reports on anomalous precursory seismicity changes [62], there is no precursor in sight which obeys a degree of universality that would make it practically useful. It is, therefore, important to study less frequent precursory phenomena by means of long model simulations.

Toward this goal, we discuss a numerical model which is on one hand reasonably physical, and on the other hand simple enough that it allows to obtain some analytical results and perform long simulations. The basic version of the model consists of a segmented two-dimensional strike-slip fault in a three-dimensional elastic half space and is inherently discrete because of the abrupt transition from static to kinetic friction [10]. This paper and ► Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of summarize a large body of analytical and numerical results associated with the model.

The results of the simulations indicate an overall good agreement of the synthetic seismicity with natural earthquake activity, with respect to frequency-size distributions and various features of earthquake sequences. The degree of spatial heterogeneity on the fault, which is implemented by means of space-dependent rheological properties, has important effects on the resulting catalogs. Smooth faults are associated with the characteristic earthquake statistics, regular occurrence of mainshocks and overall smooth stress fields. On the other hand, rough faults generate scale-free Gutenberg–Richter statistics, irregular mainshock occurrence, and overall rough stress fields. A closer look at the disorder of the stress field shows, however, that even on a smooth fault a gradual roughening takes place when the next large earthquake is approached [12,82]. This is reflected in the frequency-size distribution which evolves towards the Gutenberg–Richter law and other changes of seismicity. The results can be used to establish relations between the proximity of a state on a fault to a critical point, the (unobservable) stress field, and the (observable) seismicity functions. Furthermore, it is demonstrated that the concept of "self-organized criticality" can be folded back to criticality associated with tuning parameters [12,31]. We note that phase diagrams with different dynamic regimes as functions of tuning parameters, in addition to criticality, provide a general and rich description of seismicity. Accelerating seismic release, growing spatial correlation length, changes of frequency-size statistics and evolution of other seismicity parameters may be used to track the approach to criticality [73,75,76,77].

Future Directions

We have demonstrated that numerical fault models are valuable for understanding the underlying mechanisms of observed seismicity patterns, as well as for practical estimates of future seismic hazard. The latter requires model realizations that are tuned to a specific fault zone by assimilating available observational results and their uncertainties. In a case study, the seismic hazard in the Parkfield region has been estimated by combining such a tuned model with few observational data. The use of Bayesian analysis allows us to construct a flexible hazard model for this region which can, in general, incorporate statistical and non-statistical data (e. g. from paleoseismology and geodesy) to improve and update the estimations of the seismic hazard. This approach is particularly promising for less-well monitored regions, and especially for low-seismicity regions like those in central Europe.

Modification of the stress transfer calculations to account for a statistical preference of earthquake propagation direction on a given fault section, e. g. [6,25], can improve the estimates of seismic hazard associated with large faults. It is also possible to extend the discussed framework to other geohazards with even smaller amount of observational data, e. g. the occurrence of landslides. Since the fault model deals with coupled physical processes leading to interacting earthquakes, a challenging future direction will be the design of a more general model for interacting geohazards including earthquakes on different faults as well as landslides triggered by earthquakes, and perhaps tsunamis initiated by (submarine) earthquakes or landslides.

Acknowledgments

We thank Jürgen Kurths, James R. Rice, Frank Scherbaum, Karin Dahmen, Donald L. Turcotte, and many others for useful discussions. GZ and MH acknowledge support from the collaborative research center "Complex Nonlinear Processes" (SFB 555) of the German Research Society (DFG). SH acknowledges support from the DFG-project SCHE280/14 and the EU-project SAFER. YBZ acknowledges support from the National Science Foundation, the United States Geological Survey, and the Southern California Earthquake Center. GZ and YBZ thank the Kavli Institute for Theoretical Physics, UC Santa Barbara, for hospitality during a 2005 program on Friction, Fracture and Earthquake Physics, and partial support based on NSF grant PHY99–0794. We thank James R. Holliday, Vladimir Keilis-Borok and Willie Lee for providing comments on the paper.

Bibliography

1. Aki K, Richards PG (2002) Quantitative seismology. University Science Books, Sansalito
2. Bak P (1996) How Nature Works. The science of self-organised criticality. Springer, New York

3. Bak P, Tang C (1989) Earthquakes as a phenomenon of self-organised criticality. J Geophys Res 94:15635–156637
4. Båth M (1965) Lateral inhomogeneities in the upper mantle. Tectonophysics 2:483–514
5. Ben-Zion Y (1996) Stress, slip, and earthquakes in models of complex single-fault systems incorporating brittle and creep deformations. J Geophys Res 101:5677–5706
6. Ben-Zion Y (2001) Dynamic rupture in recent models of earthquake faults. J Mech Phys Solids 49:2209–2244
7. Ben-Zion Y (2003) Appendix 2, Key Formulas in Earthquake Seismology. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part B. Academic Press, San Diego, pp 1857–1875
8. Ben-Zion Y, Lyakhovsky V (2002) Accelerated seismic release and related aspects of seismicity patterns on earthquake faults. Pure Appl Geophys 159:2385–2412
9. Ben-Zion Y, Lyakhovsky V (2006) Analysis of aftershocks in a lithospheric model with seismogenic zone governed by damage rheology. J Geophys Int 165:197–210; doi:10.1111/j.1365-246X2006.02878.x
10. Ben-Zion Y, Rice JR (1993) Earthquake failure sequences along a cellular fault zone in a three-dimensional elastic solid containing asperity and nonasperity regions. J Geophys Res 98:14109–14131
11. Ben-Zion Y, Dahmen K, Lyakhovsky V, Ertas D, Agnon A (1999) Self-driven mode switching of earthquake activity on a fault system. Earth Plan Sci Lett 172:11–21
12. Ben-Zion Y, Eneva M, Liu Y (2003) Large Earthquake Cycles and Intermittent Criticality On Heterogeneous Faults Due To Evolving Stress and Seismicity. J Geophys Res 108:2307; doi:10.1029/2002JB002121
13. Bernardo JM, Smith AFM (1994) Bayesian Theory. Wiley, Chichester
14. Binney JJ, Dowrick NJ, Fisher AJ, Newman MEJ (1993) The theory of critical phenomena. Oxford University Press, Oxford
15. Bowman DD, Oullion G, Sammis CG, Sornette A, Sornette D (1998) An observational test of the critical earthquake concept. J Geophys Res 103:24359–24372
16. Brace WF (1960) An extension of the Griffith theory of fracture to rocks. J Geophys Res 65:3477–3480
17. Bufe CG, Varnes DJ (1993) Predicitive modeling of the seismic cycle of the greater San Francisco Bay region. J Geophys Res 98:9871–9883
18. Burridge R, Knopoff L (1967) Model and theoretical seismicity. Bull Seim Soc Am 57:341–371
19. Byerlee JD (1978) Friction of rocks. Pure Appl Geophys 116:615–616
20. Chinnery M (1963) The stress changes that accompany strike-slip faulting. Bull Seim Soc Am 53:921–932
21. Corral Á (2004) Long-term clustering, scaling, and universality in the temporal occurrence of earthquakes. Phys Rev Lett 92:108501; doi:10.1103/PhysRevLett.92.108501
22. Dahmen K, Ertas D, Ben-Zion Y (1998) Gutenberg–Richter and characteristic earthquake behavior in simple mean-field models of heterogeneous faults. Phys Rev E 58:1494–1501
23. Daley DJ, Vere-Jones D (1988) An Introduction to the Theory of Point Processes, Springer Series: Probability and its Applications. Springer, Heidelberg
24. Dieterich JH (1994) A constitutive law for earthquake production and its application to earthquake clustering. J Geophys Res 99:2601–2618
25. Dor O, Rockwell TK, Ben-Zion Y (2006) Geologic observations of damage asymmetry in the structure of the San Jacinto, San Andreas and Punchbowl faults in southern California: A possible indicator for preferred rupture propagation direction. Pure Appl Geophys 163:301–349; doi:10.1007/s00024-005-0023-9
26. Ellsworth WL, Matthews MV, Nadeau RM, Nishenko SP, Reasenberg PA, Simpson RW (1999) A physically based earthquake recurrence model for estimation of long-term earthquake probabilities. US Geol Surv Open-File Rept, pp 99–522
27. Fisher DS, Dahmen K, Ramanathan S, Ben-Zion Y (1997) Statistics of earthquakes in simple models of heterogeneous faults. Phys Rev Lett 78:4885–4888
28. Geller RJ, Jackson DD, Kagan YY, Mulargia F (1997) Earthquakes cannot be predicted. Science 275:1616–1617
29. Gumbel EJ (1960) Multivariate Extremal Distributions. Bull Inst Int Stat 37:471–475
30. Gutenberg B, Richter CF (1956) Earthquake magnitude, intensity, energy and acceleration. Bull Seismol Soc Am 46:105–145
31. Hainzl S, Zöller G (2001) The role of disorder and stress concentration in nonconservative fault systems. Phys A 294:67–84
32. Hainzl S, Zöller G, Kurths J (1999) Similar power laws for fore- and aftershock sequences in a spring-block model for earthquakes. J Geophys Res 104:7243–7253
33. Hainzl S, Zöller G, Kurths J (2000) Self-organization of spatio-temporal earthquake clusters. Nonlin Proc Geophys 7:21–29
34. Hainzl S, Zöller G, Kurths J, Zschau J (2000) Seismic quiescence as an indicator for large earthquakes in a system of self-organized criticality. Geophys Res Lett 27:597–600
35. Hainzl S, Zöller G, Scherbaum F (2003) Earthquake clusters resulting from delayed rupture propagation in finit fault segments. J Geophys Res 108:2013; doi:10.1029/2001JB000610
36. Hillers G, Mai PM, Ben-Zion Y, Ampuero JP (2007) Statistical Properties of Seismicity Along Fault Zones at Different Evolutionary Stages. J Geophys Int 169:515–533; doi:10.1111/j.1365-246X2006.03275.x
37. Huang J, Turcotte DL (1990) Are earthquakes an example of deterministic chaos? Geophys Res Lett 17:223–226
38. Jaumé SC, Sykes LR (1999) Evolving towards a critical point: A review of accelerating seismic moment/energy release prior to large and great earthquakes. Pure Appl Geophys 155:279–306
39. Jones LM, Molnar P (1979) Some characteristics of foreshocks and their possible relation to earthquake prediction and premonitory slip on faults. J Geophys Res 84:3596–3608
40. Kagan YY, Knopoff L (1978) Statistical study of the occurrence of shallow earthquakes. J Geophys R Astron Soc 55:67–86
41. Keilis-Borok VI, Soloviev AA (2003) Nonlinear Dynamics of the Lithosphere and Earthquake Prediction, Springer Series in Synergetics. Springer, Heidelberg
42. Lomnitz-Adler J (1999) Automaton models of seismic fracture: constraints imposed by the magnitude-frequency relation. J Geophys Res 98:17745–17756
43. Main IG, O'Brian G, Henderson JR (2000) Statistical physics of earthquakes: Comparison of distribution exponents for source area and potential energy and the dynamic emergence of log-periodic quanta. J Geophys Res 105:6105–6126

44. Matthews MV, Ellsworth WL, Reasenberg PA (2002) A Brownian model for recurrent earthquakes. Bull Seism Soc Am 92:2233–2250
45. Narteau C, Shebalin P, Hainzl S, Zöller G, Holschneider M (2003) Emergence of a band-limited power law in the aftershock decay rate of a slider-block model. Geophys Res Lett 30:1568; doi:10.1029/2003GL017110
46. Nur A, Booker JR (1972) Aftershocks caused by pore fluid flow? Science 175:885–887
47. Okada Y (1992) Internal deformation due to shear and tensile faults in a half space. Bull Seism Soc Am 82:1018–1040
48. Olami Z, Feder HS, Christensen K (1992) Self-organized criticality in a continuous, nonconservative cellular automaton modeling earthquakes. Phys Rev Lett 68:1244–1247
49. Omori F (1894) On the aftershocks of earthquakes. J Coll Sci Imp Univ Tokyo 7:111–200
50. Patel JK, Kapadia CH, Owen DB (1976) Handbook of statistical distributions. Marcel Dekker, New York
51. Reasenberg P (1985) Second-order moment of central California seismicity. J Geophys Res 90:5479–5495
52. Reid HF (1910) The Mechanics of the Earthquake, The California Earthquake of April 18, 1906. Report of the State Investigation Commission, vol 2. Carnegie Institution of Washington, Washington
53. Rundle JB, Klein W, Turcotte DL, Malamud BD (2000) Precursory seismic activation and critical point phenomena. Pure Appl Geophys 157:2165–2182
54. Saleur H, Sammis CG, Sornette D (1996) Discrete scale invariance, complex fractal dimensions, and log-periodic fluctuations in seismicity. J Geophys Res 101:17661–17677
55. Savage JC, Svarc JL, Prescott WH (1999) Geodetic estimates of fault slip rates in the San Francisco Bay area. J Geophys Res 104:4995–5002
56. Scholz CH (1998) Earthquakes and friction laws. Nature 391:37–42
57. Shcherbakov R, Turcotte DL (2004) A damage mechanics model for aftershocks. Pure Appl Geophys 161:2379; doi:10.1007/s00024-004-2570-x
58. Shin TC, Teng TL (2001) An overview of the 1999, Chichi, Taiwan, earthquake. Bull Seismol Soc Am 91:895–913
59. Sornette D (2004) Self-organization and Disorder: Concepts & Tools, Springer Series in Synergetics. Springer, Heidelberg
60. Sornette D, Sammis CG (1995) Complex critical exponents from renormalization group theory of earthquakes: Implication for earthquake predicitions. J Phys 1(5):607–619
61. Sornette D, Sornette A (1999) Renormalization of earthquake aftershocks. Geophys Res Lett 6:1981–1984
62. Field EH et al. (2007) Special Issue: Regional Earthquake Likelihood Models. Seismol Res Lett 78:1
63. Steacy SJ, McCloskey J, Bean CJ, Ren JW (1996) Heterogeneity in a self-organized critical earthquake model. Geophys Res Lett 23:383–386
64. Turcotte DL (1997) Fractals and chaos in geology and geophysics. Cambridge University Press, New York
65. Turcotte DL, Newman WI, Shcherbakov R (2003) Micro and macroscopic models of rock fracture. J Geophys Int 152:718–728
66. Utsu T (2002) Statistical features of seismicity. In: Int Assoc Seismol & Phys Earth's Interior (ed) International handbook of earthquake and engineering seismology, vol 81A. Academic Press, San Diego, pp 719–732
67. Utsu T, Ogata Y, Matsu'ura RS (1995) The centenary of the Omori formula for a decay law of aftershock activity. J Phys Earth 43:1–33
68. Wesnousky SG (1994) The Gutenberg–Richter or characteristic earthquake distribution, which is it? Bull Seismol Soc Am 90:525–530; 84:1940–1959
69. Wiemer S, Baer M (2000) Mapping and removing quarry blast events from seismic catalogs: Examples from Alaska, the Western United States, and Japan. Bull Seismol Soc Am 90:525–530
70. Working Group on California Earthquake Probabilities (2003) Earthquake probabilities in the San Francisco Bay region. US Geol Survey Open File Report 03–214, US Geological Survey
71. Wyss M (1997) Cannot earthquakes be predicted? Science 278:487
72. Wyss M, Habermann RE (1988) Precursory seismic quiescence. Pure Appl Geophys 126:319–332
73. Zaliapin I, Liu Z, Zöller G, Keilis-Borok V, Turcotte DL (2002) On increase of earthquake correlation length prior to large earthquakes in California. Comp Seismol 33:141–161
74. Zöller G, Ben-Zion Y, Holschneider M, Hainzl S (2007) Estimating recurrence times and seismic hazard of large earthquakes on an individual fault. J Geophys Int 170:1300–1310; doi:10.1111/j.1365-246X200703480.x
75. Zöller G, Hainzl S (2001) Detecting premonitory seismicity patterns based on critical point dynamics. Nat Hazards Earth Syst Sci 1:93–98
76. Zöller G, Hainzl S (2002) A systematic spatiotemporal test of the critical point hypothesis for large earthquakes. Geophys Res Lett 29:1558; doi:10.1029/2002GL014856
77. Zöller G, Hainzl S, Kurths J (2001) Observation of growing correlation length as an indicator for critical point behavior prior to large earthquakes. J Geophys Res 106:2167–2175
78. Zöller G, Hainzl S, Kurths J, Zschau J (2002) A systematic test on precursory seismic quiescence in Armenia. Nat Hazards 26:245–263
79. Zöller G, Holschneider M, Ben-Zion Y (2004) Quasi-static and quasi-dynamic modeling of earthquake failure at intermediate scales. Pure Appl Geophys 161:2103–2118; doi:10.1007/s00024-004-2551-0
80. Zöller G, Holschneider M, Ben-Zion Y (2005) The role of heterogeneities as a tuning parameter of earthquake dynamics. Pure Appl Geophys 162:1027; doi:10.1007/s00024-004-2660-9
81. Zöller G, Hainzl S, Holschneider M, Ben-Zion Y (2005) Aftershocks resulting from creeping sections in a heterogeneous fault. Geophys Res Lett 32:L03308; doi:10.1029/2004GL021871
82. Zöller G, Hainzl S, Ben-Zion Y, Holschneider M (2006) Earthquake activity related to seismic cycles in a model for a heterogeneous strike-slip fault. Tectonophys 423:137–145; doi:10.1016/j.tecto.2006.03.007

Seismicity, Statistical Physics Approaches to

DIDIER SORNETTE[1], MAXIMILIAN J. WERNER[2]

[1] Department of Management, Technology
and Economics, ETH Zurich, Switzerland

[2] Swiss Seismological Service, Institute of Geophysics,
ETH Zurich, Switzerland

Article Outline

Glossary

Chaos Chaos occurs in dynamical systems with two ingredients: (i) nonlinear recurrent re-injection of the dynamics into a finite domain in phase space and (ii) exponential sensitivity of the trajectories in phase space to initial conditions.

Continuous phase transitions If there is a finite discontinuity in the first derivative of the thermodynamic potential, then the phase transition is termed first-order. During such a transition, a system either absorbs or releases a fixed amount of latent heat (e. g. the freezing/melting of water/ice). If the first derivative is continuous but higher derivatives are discontinuous or infinite, then the phase transition is called continuous, of the second kind, or critical. Examples include the critical point of the liquid–gas transition, the Curie point of the ferromagnetic transition, or the superfluid transition [127,235].

Critical exponents Near the critical point, various thermodynamic quantities diverge as power laws with associated critical exponents. In equilibrium systems, there are scaling relations that connect some of the critical exponents of different thermodynamic quantities [32,127,203,216,235].

Critical phenomena Phenomena observed in systems that undergo a continuous phase transition. They are characterized by scale invariance: the statistical properties of a system at one scale are related to those at another scale only through the ratio of the two scales and not through any one of the two scales individually. The scale invariance is a result of fluctuations and correlations at all scales, which prevents the system from being separable in the large scale limit at the critical point [32,203,235].

Declustering In studies of seismicity, declustering traditionally refers to the deterministic identification of fore-, main- and aftershocks in sequences (or clusters) of earthquakes clustered in time and space. Recent, more sophisticated techniques, e. g. stochastic declustering, assign to earthquakes probabilities of being triggered or spontaneous.

Dynamical scaling and exponents Non-equilibrium critical phase transitions are also characterized by scale invariance, scaling functions and critical exponents. Furthermore, some evidence supports the claim that universality classes also exist for non-equilibrium phase transitions (e. g. the directed percolation and the Manna universality class in sandpile models), although a complete classification of classes is lacking and may in fact not exist at all. Much interest has recently focused on directed percolation, which, as the most common universality class of absorbing state phase transitions, is expected to occur in many physical, chemical and biological systems [85,135,203].

Finite size scaling If a thermodynamic or other quantity is investigated at the critical point under a change of the system size, the scaling behavior of the quantity with respect to the system size is known as finite size scaling [32]. The quantity may refer to a thermodynamic quantity such as the free energy or it may refer to an entire probability distribution function. At criticality, the sole length scale in a finite system is the upper cut-off s_c, which diverges in the thermodynamic limit $L \to \infty$. Assuming a lower cut-off $s_0 \ll s_c, s$, a finite size scaling ansatz for the distribution $P(s; s_c)$ of the observable variable s, which depends on the upper cut-off s_c is then given by:

$$P(s; s_c) = a s^{-\tau} G(s/s_c) \quad \text{for} \quad s, s_c \gg s_0 \,, \tag{1}$$

where the parameter a is a non-universal metric factor, τ is a universal (critical) exponent, and G is a universal scaling function that decays sufficiently fast for $s \gg s_c$ [32,36]. Pruessner [163] provides a simple yet instructive and concise introduction to scaling theory and how to find associated exponents. System-specific corrections appear to sub-leading order.

Fractal A deterministic or stochastic mathematical object that is defined by its exact or statistical self-similarity at all scales. Informally, it often refers to a rough or fragmented geometrical shape which can be subdivided into parts which look approximately the same as the original shape. A fractal is too irregular to be described by Euclidean geometry and has a fractal dimension that is larger than its topological dimension but less than the dimension of the space it occupies.

Mean-Field An effective or average interaction field designed to approximately replace the interactions from many bodies by one effective interaction which is constant in time and space, neglecting fluctuations.

Mechanisms for power laws Power laws may be the hallmark of critical phenomena, but there are a host of other mechanisms that can lead to power laws (see Chapter 14 of [203] for a list of power law mechanisms as well as [37,143]). Observations of scale invariant statistics therefore do not necessarily imply SOC, of course. Power laws express the existence of a symmetry (scale invariance) and there are many mechanisms by which a symmetry can be obtained or restored.

Non-equilibrium phase transitions In contrast to systems at equilibrium, non-equilibrium phase transitions involve dynamics, energy input and dissipation. Detailed balance is violated and no known equivalent of the partition function exists, from which all thermodynamic quantities of interest derive in equilibrium. Examples of non-equilibrium phase transitions include absorbing state phase transitions, reaction-diffusion models, and morphological transitions of growing surfaces [85,135].

Phase transitions In (equilibrium) statistical mechanics, a phase transition occurs when there is a singularity in the free energy or one of its derivatives. Examples include the freezing of water, the transition from ferromagnetic to paramagnetic behavior in magnets, and the transition from a normal conductor to a superconductor [127,235].

Renormalization group theory A mathematical theory built on the idea that the critical point can be mapped onto a fixed point of a suitably chosen transformation on the system's Hamiltonian. It provides a foundation for understanding scaling and universality and provides tools for calculating exponents and scaling functions. Renormalization group theory provides the basis for our understanding of critical phenomena [32,216,235]. It has been extended to non-Hamiltonian systems and provides a general framework for constructing theories of the macro-world from the microscopic description.

Self-organized criticality (SOC) Despite two decades of research since its inception by [13] and the ambitious claim by [11] that, as a mechanism for the ubiquitous power laws in Nature, SOC was "How Nature Works", a commonly accepted definition along with necessary and sufficient conditions for SOC is still lacking [93,163,203]. A less rigorous definition may be the following: Self-organized criticality refers to a non-equilibrium, critical and marginally stable steady-state, which is attained spontaneously and without (explicit) tuning of parameters. It is characterized by power law event distributions and fractal geometry (in some cases) and may be expected in slowly driven, interaction-dominated threshold systems [93]. Some authors additionally require that temporal and/or spatial correlations decay algebraically (e. g. [84], but see [163]). Definitions in the literature range from broad (simply the absence of characteristic length scales in non-equilibrium systems) to narrow (the criticality is due to an underlying continuous phase transition with all of its expected properties) (see, e. g., [162] for evidence that precipitation is an instance of the latter definition of SOC in which a non-linear feedback of the order parameter on the control parameter turns a critical phase transition into a self-organized one attracting the dynamics [198]).

Spinodal decomposition In contrast to the slow process of phase separation via nucleation and slow growth of a new phase in a material inside the metasstable region near a first-order phase transition, spinodal decomposition is a non-equilibrium, rapid and critical-like dynamical process of phase separation that occurs quickly and throughout the material. It needs to be induced by rapidly quenching the material to reach a sub-area (sometimes a line) of the unstable region of the phase diagram which is characterized by a negative derivative of the free energy.

Statistical physics is the set of concepts and mathematical techniques allowing one to derive the large-scale laws of a physical system from the specification of the relevant microscopic elements and of their interactions.

Turbulence In fluid mechanics, turbulence refers to a regime in which the dynamics of the flow involves many interacting degrees of freedom, and is very complex with intermittent velocity bursts leading to anomalous scaling laws describing the energy transfer from injection at large scales to dissipation at small scales.

Universality In systems with little or no frozen disorder, equilibrium continuous phase transitions fall into a small set of universality classes that are characterized by the same critical exponents and by certain scaling functions that become identical near the critical point. The class depends only on the dimension of the space and the dimension of the order parameter. For instance, the critical point of the liquid–gas transition falls into the same universality class as the 3D Ising model. Even some phase transitions occurring in high-energy physics are expected to belong to the Ising class. Universality justifies the development and study of extremely simplified models (caricatures) of Nature, since the behavior of the system at the critical point can nevertheless be captured (in some cases exactly). However, non-universal features remain even at the critical point but are less important, e. g. amplitudes of fluctuations or system-specific corrections to scaling that appear at sub-leading order [32,216,235,239].

Definition of the Subject

A fundamental challenge in many scientific disciplines concerns upscaling, that is, of determining the regularities and laws of evolution at some large scale from those known at a lower scale: biology (from molecules to cells, from cells to organs); neurobiology (from neurons to brain function), psychology (from brain to emotions, from evolution to understanding), ecology (from species to the global web of ecological interactions), condensed matter physics (from atoms and molecules to organized phases such as solid, liquid, gas, and intermediate structures), social sciences (from individual humans to social groups and to society), economics (from producers and consumers to the whole economy), finance (from investors to the global financial markets), Internet (from e-pages to the world wide web 2.0), semantics (from letters and words to sentences and meaning), and so on. Earthquake physics is no exception, with the challenge of understanding the transition from the laboratory scale (or even the microscopic and atomic scale) to the scale of fault networks and large earthquakes.

Statistical physics has had a remarkably successful track record in addressing the upscaling problem in physics. While the macroscopic laws of thermodynamics have been established in the 19th century, their microscopic underpinning were elaborated in the early 20th century by Boltzmann and followers, building the magnificent edifice of statistical physics. Statistical physics can be defined as the set of concepts and mathematical techniques allowing one to derive the large-scale laws of a physical system from the specification of the relevant microscopic elements and of their interactions. Dealing with huge ensembles of elements (atoms, molecules) of the order of the Avogadro number ($\simeq 6 \cdot 10^{23}$), statistical physics uses the mathematical tools of probability theory combined with other relevant fields of physics to calculate the macroscopic properties of large populations.

One of the greatest achievement of statistical physics was the development of the renormalization group analysis, to construct a theory of interacting fields and of critical phase transitions. The renormalization group is a perfect example of how statistical physics addresses the micro-macro upscaling problem. It decomposes a problem of finding the macroscopic behavior of a large number of interacting parts into a succession of simpler problems with a decreasing number of interacting parts, whose effective properties vary with the scale of observation. The renormalization group thus follows the proverb "divide to conquer" by organizing the description of a system scale-by-scale. It is particularly adapted to critical phenomena and to systems close to being scale-invariant. The renormalization group translates into mathematical language the concept that the overall behavior of a system is the aggregation of an ensemble of arbitrarily defined sub-systems, with each sub-system defined by the aggregation of sub-subsystems, and so on [203].

It is important to stress that up to now the term "statistical" has different meanings in statistical physics and in statistical seismology, a field that has developed as a marriage between probability theory, statistics and the part of seismology concerned with empirical patterns of earthquake occurrences [225] (but not with physics). Statistical seismology uses stochastic models of seismicity, which are already effective large-scale representation of the dynamical organization. In contrast, a statistical physics approach to earthquake strives to derive these statistical models or other descriptions from the knowledge of the microscopic laws of friction, damage, rupture, rock-water interactions, mechano-chemistry and so on, at the microscopic scales [200,201]. In other words, what is often missing in statistical seismology is the physics to underpin the stochastic model on physically-based laws, e. g. rate-and-state friction [55].

The previously mentioned successes of statistical physics promote the hope that a similar program can be developed for other fields, including seismology. The successes have been more limited, due to the much more complex interplay between mechanisms, interactions and scales found in these out-of-equilibrium systems. This short essay provides a subjective entry to understand some of the different attempts, underlining the few successes, the problems and open questions. Rather than providing an exhaustive review, we mention what we believe to be important topics and have especially included recent work.

Introduction

Much of the recent interest of the statistical physics community has focused on applying scaling techniques, which are common tools in the study of critical phenomena, to the statistics of inter-event recurrence times or waiting times [14,40,41,42,43,44,46,48,134]. However, the debate over the relevance of critical phenomena to earthquakes stretches back as far as 30 years [7,11,12,47,61,84,93,104,106,109,114,147,151,178,192,193,194,202,203,207,223], ► Jerky Motion in

Slowly Driven Magnetic and Earthquake Fault Systems, Physics of and ► Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space. The current debate on recurrence statistics is thus the latest tack in an evolving string of arguments with a long history. As discussed below, many of the claims made in the recent articles on recurrence statistics have either been challenged, refuted or explained by previously known facts about earthquake statistics [132,133,145,176,177,230]. As will be discussed below, this debate in the literature is important because of the potential consequences for understanding earthquakes, but it needs to be pursued with rigorous scientific arguments accessible to both the seismological and the statistical physics communities.

The debate would almost certainly benefit significantly from testing hypotheses with simulations to establish null hypotheses and benchmarks: seismicity patterns are sufficiently stochastic and earthquake catalogs contain a sufficient amount of observational uncertainties so as to make inference difficult. It is often not straightforward to predict the signal of well-known statistical features such as clustering in new data analysis techniques. Therefore, testing the purported claims by realistic simulations of earthquake catalogs can provide a strong benchmark against which the claims can be evaluated. This view and the corresponding criticism of many studies has been put forward and defended for a long time by Kagan [111].

Such a model-dependent approach may be at odds with the philosophy of a so-called "model-free" analysis, which the community of statistical physicists claim to take in their analysis. For instance, network theory-based approaches, space-time window-based finite size scaling, box-covering methods and other techniques used in the study of critical and fractal phenomena are said to be "model-free" because no assumptions about seismicity are supposedly made at the outset. By using model-free analysis techniques, the often uncertain and sometimes clearly wrong assumptions of flawed models and resulting biased results are meant to be circumvented.

However, as is almost always the case in statistical hypothesis testing, the less assumptions are made about the test, the less powerful the test statistic. More importantly, seismicity is sufficiently stochastic so that well-known features may appear as novel in new analysis methods. Furthermore, to convince the seismological community of new data analysis techniques, the methods need to be tested on established knowledge and show the improvement over traditional methods. These types of initial tests are rarely performed by the statistical physics community.

In the next Sect. "Concepts and Calculational Tools", we present a summary of some of the concepts and calculational tools that have been developed in attempts to apply statistical physics approaches to seismology. Then, Sect. "Competing Mechanisms and Models" summarizes the leading theoretical physical models of the space-time organization of earthquakes. Section "Empirical Studies of Seismicity Inspired by Statistical Physics" presents a general discussion and several examples of the new metrics proposed by statistical physicists, underlining their strengths and weaknesses. Section "Future Directions" briefly outlines expected developments.

Concepts and Calculational Tools

Renormalization, Scaling and the Role of Small Earthquakes in Models of Triggered Seismicity

A common theme in many of the empirical relations in seismology (and in those employed in seismicity models) is the lack of a dominating scale. Many natural phenomena can be approached by the traditional reductionist approach to isolate a process at a particular scale. For example, the waves of an ocean can be described quite accurately by a theory that entirely ignores the fact that the liquid is made out of individual molecules. Indeed, the success of most practical theories in physics depends on isolating a scale [234], although since this recognition, much progress has been made in developing a holistic approach for processes that do not fall into this class. Given current observational evidence, earthquakes seem to belong to the set of processes characterized by a lack of one dominating length scale: fluctuations of many or perhaps a wide continuum of sizes seem to be important and are in no way diminished – even when one is interested solely in large-scale descriptions [206].

The traditional reductionist approach in seismology, which, for instance, attempted to separate large (main) shocks from small (fore- or after-) shocks, is slowly giving way to the holistic approach, in which all earthquakes are created equal and seismicity is characterized by fluctuations of all sizes. This gradual shift is supported, on a conceptual and qualitative level, by the vision of critical phenomena. A particularly strong model of the interactions between earthquakes has emerged in the concept of triggering, which places all earthquakes on the same footing: each earthquake can trigger its own events, which in turn can trigger their own events, and so on, according to the same probability distributions, and the resulting seismicity can be viewed as the superposed cascades of triggered earthquakes that cluster in space and time [77,117,149,150].

From this point of view, it is natural that small earthquakes are important to the overall spatio-temporal patterns of seismicity. Indeed, the scaling of aftershock productivity with mainshock magnitude suggests that small earthquakes are cumulatively as important for the triggered seismicity budget as rarer but larger events [60,76,82]. The importance of small earthquakes has also been documented in, e. g., [73,138,141].

But earthquake catalogs do not contain information (by definition) about the smallest, unobserved events, which we know to exist from acoustic emission experiments and earthquakes recorded in mines. To guarantee a finite seismicity budget, Sornette and Werner argued [209] for the existence of a smallest triggering earthquake, akin to a "ultra-violet cut-off" in quantum field theory, below which earthquakes do not trigger other events. Introducing a formalism which distinguishes between the detection threshold and the smallest triggering earthquake, Sornette and Werner placed constraints on its size by using a simplified version of the popular Epidemic-Type Aftershock Sequence (ETAS) Model [149], a powerful model of triggered seismicity based on empirical statistics, and by using observed aftershock sequences. Sornette and Werner [210] also considered the branching structure of one complete cascade of triggered events, deriving an apparent branching ratio and the apparent number of untriggered events, which are observed when only the structure above the detection threshold is known. As a result of our inability to observe the entire branching structure, inferred clustering parameters are significantly biased and difficult to interpret in geophysical terms. Second, separating triggered from untriggered events, commonly known as declustering, also strongly depends on the threshold, so that it cannot even in theory constitute a physically sound method.

Sornette and Werner [210] also found that a simplified, averaged version of the ETAS model can be renormalized onto itself, with effective clustering parameters, under a change of the threshold. Saichev and Sornette [175] confirmed these results for the stochastic number statistics of the model using a rigorous approach in terms of generating probability functions, but also showed that the temporal statistics could not be renormalized. Furthermore, it can be shown (see Chapter 4 of [229]) that the conditional intensity function of the ETAS model, the mathematical object which uniquely defines the model, cannot be renormalized onto itself under a change of magnitude threshold. It is not a fixed-point of the renormalization process operating via magnitude coarse-graining. The functional form of the model must change under a change in the detection threshold [175]. In other words, if earthquakes occur according to an ETAS model above some cut-off m_0, then earthquakes above m_d cannot be described by the ETAS model in a mathematically exact way. Although in practice, the ETAS model provides an excellent fit.

The issue of how to deal with small earthquakes is thus reminiscent of the decades of efforts that have been invested in physics to deal with the famous ultra-violet cut-off problem, eventually solved by the so-called "renormalization" theory of Feynmann, Schwinger and Tomonaga. In the 1960s and 1970s, this method of renormalization was extended into the "renormalization group" (in fact a semi-group in the strict mathematical sense) for the theory of critical phenomena (see glossary), which we also mention in Sect. "Competing Mechanisms and Models". It is fair to say that there has been limited success in developing a multi-scale description of the physics of earthquakes and, in particular, in addressing the upscaling problem and the impact of the many small earthquakes.

One tantalizing approach, not yet really understood in terms of all its consequences and predictions, is the variant of the ETAS model proposed by Vere-Jones [224], which has the remarkable property of being bi-scale invariant under a transformation involving time and magnitudes. One of the modifications brought in by [224] is to assume that the distribution of the daughter magnitudes is dependent on the mother magnitude m_i through a modification of the Gutenberg–Richter distribution of triggered earthquake magnitudes by a term of the form $\exp(-\delta|m - m_i|)$, where $\delta > 0$ quantifies the distance to the standard Gutenberg–Richter distribution. Remarkably, Saichev and Sornette [174], who studied the Vere-Jones model, found that, due to the superposition of the many magnitude distributions of each earthquake in the cascades of triggered events, the resulting distribution of magnitudes over a stationary catalog is a pure Gutenberg–Richter law. Thus, there might be hidden characteristic scales in the physics of triggering that are not revealed by the standard observable one-point statistical distributions. Simulation and parameter estimation algorithms for the Vere-Jones model are not yet available. If and when these algorithms become available, the study of this bi-scale invariant branching model may be a strong alternative to the ETAS model, as this model is exactly scale invariant with neither ultra-violet nor infra-red cut-offs.

Nevertheless, being empirically based, these stochastic point process models lack a genuine microscopic physical foundation. The underlying physics is not explicitly addressed and only captured effectively by empirical statistics, even at the smallest scales. The physical processes and their renormalization are missing in this approach.

Universality

Universality, as defined in the glossary, justifies the development and study of extremely simplified models (caricatures) of Nature, since the behavior of a studied system at the critical point can nevertheless be captured by toy models (in some cases exactly). For instance, the liquid–gas transition, the ferromagnetic to paramagnetic transition, and the behavior of binary alloys, all apparently different systems, can be described successfully by an extremely simplified picture of Nature (the Ising model) because of universality. The hope that a similar principle holds for earthquakes (and other non-equilbrium systems) underpins many of the models and tools inspired by statistical physics that have been applied to seismicity, to bring about the "much coveted revolution beyond reductionism" [17,67].

Speaking loosely, the appearance of power laws in many toy models is often interpreted as a kind of universal behavior. The strict interpretation of universality classes, however, requires that critical exponents along with scaling functions are identical for different systems. It is interesting to note that slight changes in the sand-pile model already induce new universality classes, so that even within a group of toy models, the promise of universality is not, strictly, fulfilled [13,98,136].

A lively debate in seismology concerns the universality of, on one hand, the frequency-size distribution of earthquake magnitudes (e. g. [47,61,91,108,111,232]), and, on the other hand, the universality of the exponent of the Gutenberg–Richter distribution (e. g. [25,184,212,233]). A spatio-temporally varying critical exponent is not traditionally part of the standard critical phenomena repertoire, although analytical and numerical results based on a simple earthquake model on a fault showed a possible spontaneous switching between Gutenberg–Richter and characteristic earthquake behavior associated with a non-equilibrium phase transition [24,47,61], ► Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of and ► Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space. Another possible mechanism for the coexistence of and intermittent shifts between different regimes (Gutenberg–Richter scaling and characteristic earthquakes) stems from the competition between several interacting faults which give rise to long intervals of activity in some regions followed by similarly long intervals of quiescence [129,213,214]. Both careful empirical investigations of seismicity parameters and theoretical progress on heterogeneous, spatially extended critical phenomena may help elucidate the controversy.

Intermittent Periodicity and Chaos

As part of the conquest of chaos theory in the 70s and 80s [58,139], its concepts and methods were invariably also applied to seismicity. Huang and Turcotte [87,88] modeled the interaction between two faults by two sliding blocks that are driven by a plate through springs, coupled to one another via another spring and endowed with a velocity-weakening friction law. The dynamical evolution of the blocks showed chaotic behavior and period-doubling (the Feigenbaum route to chaos [58]). Huang and Turcotte [87,88] suggested that the interaction of the Parkfield segment with the southern San Andreas fault may be governed by the kind of chaotic behavior they observed in their model: the lighter block slipped quasi-periodically for several times until both slipped together. Their study explained that apparent quasi-periodicity of earthquakes on fault segments may be a result of chaotic interactions between many fault segments, thereby providing a warning that the extrapolations of quasi-periodic models are not to be trusted (e. g. [15,75] and references therein). However, it is doubtful that models with just a few degrees of freedom can go a long way towards providing deeper physical insights, or predictive tools. One needs to turn to models with a large number of degrees of freedom, for which turbulence appears as the leading paradigm of complexity.

Turbulence

A drastically different approach has been favored by Yan Kagan [104,106,109,114], who described seismicity as the "turbulence of solids" – attesting to the far greater problems in earthquake seismology than the theory of critical phenomena promises to solve. While renormalization group methods and scaling theory have contributed immensely to the study of turbulence [63], the problem of turbulence involves significant additional complications. First, loosely speaking, renormalization group theory helps predict global behavior by coarse-graining over degrees of freedom, which is essentially a bottom–up approach. In turbulence, the enstrophy acts bottom–up, but the energy cascades top down. Secondly, there is a significant spatial and topological aspect to turbulence, for instance involving topological defects, such as filament structures, which are crucial to the dynamical evolution of the system. The existence of the two cascades (top–down and bottom–up) as well as the influence of the dissipation scale all the way within the so-called inertial range makes turbulence the most important problem still unsolved in classical Physics. The importance of addressing the issue of the interplay between the top–down and bottom–up cascades in earthquake toy models has been outlined by [65,66,236,237].

The analogous problem for seismicity lies in the complex fault network, which constrains seismicity through its weak structures but also grows and evolves because of earthquakes redistributing stresses and rupturing fresh surfaces. The statistical description of this tensorial and dynamical problem is only at its beginning [64,99,100,101,102,103,107,116,118,119,142,213]. But it is likely to be a key aspect to the dynamical evolution of faults and seismicity. New physics and approaches are required to tackle the tensorial nature of the stress and strain fields and the complex topological structures of defects, from dislocations to joints and faults, as well as the many different physical processes operating from the atomic scale to the macro-scale [200,201].

Self-Organized Criticality

Self-organized criticality (SOC) refers to the spontaneous organization of a system driven from outside into a globally dynamical statistical stationary state, which is characterized by self-similar distributions of event sizes and sometimes fractal geometrical properties. SOC applies to the class of phenomena occurring in driven out-of-equilibrium systems made of many interactive components, which possess the following fundamental properties: 1) a highly non-linear behavior, 2) a very slow driving rate, 3) a globally stationary regime, characterized by stationary statistical properties, and 4) power-law distributions of event sizes and fractal geometrical properties. The crust obeys these four conditions, as first suggested by [12,193], who proposed to understand the spatio-temporal complexity of earthquakes in this perspective.

The appeal of placing the study of earthquakes in the framework of critical phenomena may be summarized as follows. Power law distributions can be understood as a result of an underlying continuous phase transition into which the crust has organized itself [197,211]. Applying the methods of renormalization group theory may help calculate exponents and scaling functions and rationalize the spatio-temporal organization of seismicity along with its highly correlated structures. For instance, Sornette and Virieux [208] provided a theoretical framework which links the scaling laws exhibited by earthquakes at short times and plate tectonic deformations at large times. Perhaps earthquakes fall into a universality class which can be solved exactly and/or investigated in toy models. Moreover, studying the detailed and highly complicated microphysics involved in earthquakes may not lead to insights about the spatio-temporal organization, because, as a critical phenomenon, the traditional approach of separating length scales to describe systems is inadequate. On the other hand, as mentioned above, there is the possibility of a hierarchy of physical processes and scales which are inter-related [156,157], for which the simplifying approach in terms of critical phenomena is likely to be insufficient.

As another reason for the importance of the topic, interesting consequences for the predictability of earthquakes might be derived, for instance by mapping earthquakes to a genuine critical point (the accelerating moment release hypothesis, e. g. [202,207]) or by mapping earthquakes to SOC (e. g. [70,147]). The latter mapping had led some to argue that earthquakes are inherently unpredictable. In the sandpile paradigm [13], there is little difference between small and large avalanches, and this led similarly to the concept that “large earthquakes are small earthquakes that did not stop,” hence their supposed lack of predictability. More than ten years after this contentious proposal, a majority of researchers, including most of the authors of this “impossibility claim,” recognize that there is some degree of predictability [83,97]. Actually, the clarifications came from investigators of SOC, who recognized that the long-term organization of sandpiles [51,52] and of toy models of earthquakes and fault networks [142,213,214] is characterized by long-range spatial and temporal correlations. Thus, large events may indeed be preceded by subtle long-range organizational structures, an idea at the basis of the accelerating moment release hypothesis. This idea is also underlying the pattern recognition method introduced by Gelfand et al. [69] and developed extensively by V. Keilis-Borok and his collaborators for earthquake predictions [122]. In addition, Huang et al. [89] showed that avalanche dynamics occurring within hierarchical geometric structures are characterized by significant precursory activity before large events; this provides a clear proof of the possible coexistence between critical-like precursors of large events and a long-term self-organized critical dynamical state.

In summary, self-organized criticality provides a general conceptual framework to articulate the search for a physical understanding of the large-scale and long-time statistical properties of the seismogenic process and of the predictability of earthquakes. Beyond this, it is of little help as many different mechanisms have been documented at the origin of SOC (see, e. g., chapter 15 in [203]). SOC is not a theory, it does not provide any specific calculation tools; it is a concept offering a broad classification of the kinds of dynamics that certain systems, including the Earth crust, seem to spontaneously select.

Competing Mechanisms and Models

It should be noted at this point that the statistical physics approach to earthquake science is not limited to SOC. Over the years, several groups have proposed to apply the concepts and tools of statistical physics to understanding the multiscale dynamics of earthquake and fault systems. Various mechanisms drawn conceptually from statistical mechanics but not necessary even limited to critical (phase transition) phenomena have been proposed and are being pursued. Such approaches include the concept of the critical point earthquake related to accelerated moment release, network theory, percolation and fiber models as models for fracture, and many more, some of which can be found in [84,203,220,221].

In this section, we outline some of the major model classes which underpin distinct views on what are the dominating mechanisms to understand earthquakes and their space-time organization.

Roots of Complexity in Seismicity: Dynamics or Heterogeneity?

The 1990s were characterized by vigorous discussions at the frontier between seismology and statistical physics aimed at understanding the origin of the observed complexity of the spatio-temporal patterns of earthquakes. The debate was centered on the question of whether space-time complexity can occur on a single homogeneous fault, solely as a result of the nonlinear dynamics [23,38,39,128,186,187,188,189], associated with the slip and velocity dependent friction law validated empirically in particular by [53,54,55,56]. Or, is the presence of quenched heterogeneity necessary [21,22,126,168]?

The rediscovery of the multi-slider-block-spring model of [31,33] led to a flurry of investigations by physicists [34,128,170], finding an enticing entry to this difficult field, in the hope of capturing the main empirical statistical properties of seismicity. It is now understood that complexity in the stress field, in co-seismic slips and in sequences of earthquakes can emerge purely from the nonlinear laws. However, heterogeneity is probably the most important factor dominating the multi-scale complex nature of earthquakes and faulting [156,157,181,182]. It is also known to control the appearance of self-organized critical behavior in a class of models relevant to the crust [191,214].

Critical Earthquakes

This section gives a brief history of the "critical earthquake" concept.

We trace the ancestor of the critical earthquake concept to Vere-Jones [223], who used a branching model to illustrate that rupture can proceed through a cascade of damage events. Allègre et al. [7] proposed what is in essence a percolation model of damage/rupture describing the state of the crust before an earthquake. They formulated the model using the language of real-space renormalization group, in order to emphasize the multi-scale nature of the underlying physics, and the incipient rupture as the approach to a critical percolation point. Their approach is actually a reformulation in the language of earthquakes of the real-space renormalization group approach to percolation developed by [165]. Chelidze [35] independently developed similar ideas. In the same spirit, Smalley et al. [192] proposed a renormalization group treatment of a multi-slider-block-spring model. Sornette and Sornette [194] took seriously the concept put forward by [7] and proposed to test it empirically by searching for the predicted critical precursors. Voight [227,228] was probably the first author to introduce the idea of a time-to-failure analysis quantified by a second order nonlinear ordinary differential equation. For certain values of the parameters, the solution of [227,228]'s time-to-failure equation takes the form of a finite time singularity (see [180] for a review and [204] for a mechanism based on the ETAS model). He proposed and did use it later to predict volcanic eruptions. The concept that earthquakes are somehow associated with critical phenomena was also underlying the research efforts of a part of the Russian school [120,222].

The empirical seed for the critical earthquake concept were the repeated observations that large earthquakes are sometimes preceded by an increase in the number of intermediate size events [29,59,92,96,121,123,131,144,164,217]. The relation between these intermediate events and the subsequent main event took a long time to be recognized because the precursory events occur over such a large area. Sykes and Jaumé [217] proposed a specific law $\sim \exp[t/\tau]$ quantifying the acceleration of seismicity prior to large earthquakes. Bufe and Varnes [30] proposed that the finite-time singularity law

$$\epsilon_{\text{Benioff}} \sim 1/(t_c - t)^m \qquad (2)$$

is a better empirical model than the exponential law. In (2), $\epsilon_{\text{Benioff}}$ is the cumulative Benioff strain, t_c is critical time of the occurrence of the target earthquake and m is a positive exponent. The fit with this law of the empirical Benioff strain calculated by summing the contribution of earthquakes in a given space-time window is supposed to provide the time t_c of the earthquake and thus constitutes a prediction. This expression (2) was justified by a mechanical model of material damage. It is important to understand that the law (2) can emerge as a consequence of a variety of mechanisms, as reviewed by [180].

One of these mechanisms has been coined the "critical earthquake" concept, first formulated by Sornette and Sammis [207], who proposed to reinterpret the formula (2) proposed by [30] and previous related works by generalizing them within the statistical physics framework. This concept views a large earthquake as a genuine critical point. Using the insight of critical points in rupture phenomena, Sornette and Sammis [207] proposed to enrich Eq. (2), now interpreted as a kind of diverging susceptibility in the sense of critical phenomena, by considering complex exponents (i. e. log-periodic corrections to scaling). These structures accommodate the possible presence of a hierarchy of characteristic scales, coexisting with power laws expressing the scale invariance associated with a critical phenomenon [199]. This was followed by several extensions [89,94,95,178]. Sornette [202] reviewed the concept of critical "ruptures" and earthquakes with application to prediction. Ike and Sornette [90] presented a simple dynamical mechanism to obtain finite-time singularities (in rupture in particular) decorated by complex exponents (log-periodicity). Bowman et al. [28, 153,154,242,243] proposed empirical tests of the critical earthquake concept. The early tests of [28] have been criticized by [74], while [226] commented on the lack of a formal statistical basis of the accelerating moment release model. This stresses the need for rigorous tests in the spirit of [166,167]. The debate is wide open, especially in view of the recent developments to improve the determination of the relevant spatio-temporal domain that should be used to perform the analyzes [26,27,124,130] (see [62] for a review).

Spinodal Decomposition

Klein, Rundle and their collaborators have suggested a mean-field approach to the multi-slider-block-spring model justified by the long-range nature of the elastic interactions between faults. This has led them to propose that the fluctuations of the strain and stress field associated with earthquakes are technically those occurring close to a spinodal line of an underlying first-order phase transition (see [125,169,172] and references therein). This conceptual view has inspired them to develop the "Pattern Informatics" technique, an empirical seismicity forecasting method based on the idea that changes in the seismicity rate are proxies for changes in the underlying stress [86, 218].

The fluctuations associated with a spinodal line are very similar to those observed in critical phenomena. It is thus very difficult if not impossible in principle to falsify this hypothesis against the critical earthquake hypothesis, since both are expected to present similar if not identical signatures. Perhaps, the appeal of the spinodal decomposition proposal has to be found at the theoretical level, from the fact that first-order phase transitions are more generic and robust than critical phenomena, for systems where heterogeneity and quenched randomness are not too large.

Dynamics, Stress Interaction and Thermal Fluctuation Effects

Fisher et al. [24], Dahmen et al. [47], Ben-Zion et al. [61] and co-workers (see the reviews by ▶ Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of and ▶ Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space) have introduced a mean-field model (resulting from a uniform long-range Green function) of a single fault, whose dynamical organization is controlled by two control parameters, ϵ which measures the dynamic stress weakening and c which is the deviation from stress conservation (due for instance to coupling with ductile layers). The point ($\epsilon = 0$; $c = 0$) is critical in the sense of a phase transition in statistical physics, with its associated scale invariant fluctuations described by power laws. Dynamic stress strengthening ($\epsilon < 0$) leads to truncated Gutenberg–Richter power laws. Dynamic stress weakening ($\epsilon > 0$) is either associated with a truncated Gutenberg–Richter power law for $c > 0$ or with characteristic earthquakes decorating a truncated power law for $c < 0$. The coexistence of a characteristic earthquake regime with a power-law regime is particularly interesting as it suggests that they are not exclusive properties but may characterize the same underlying physics under slightly different conditions. This could provide a step towards explaining the variety of empirical observations in seismology [5,110,185, 232].

Sornette et al. [214] have obtained similar conclusions using a quasi-static model in which faults grow and self-organize into optimal structures by repeated earthquakes. Depending on the value of dynamical stress drop (controlling the coupling strength between elements) relative to the amplitude of the frozen heterogeneity of the stress thresholds controlling the earthquake nucleation on each fault segment, a characteristic earthquake regime with a truncated power law is found for small heterogeneity or large stress drop while the power law SOC regime is recovered by large heterogeneity or small stress drop. The two approaches of [24,47,61] and ► Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of on the one hand and of [214] on the other hand can be reconciled conceptually by noting that the dynamic stress weakening of ► Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of controls the dynamical generated stress heterogeneity while the lack of stress conservation c controls the coupling strength. Fundamentally, the relevant control parameter is the degree of coupling between fault elements seen as threshold oscillators of relaxation versus the variance of the disorder in their spontaneous large earthquake recurrence times. Generically, power law statistics are expected to co-exist with synchronized behavior in a general phase diagram in the heterogeneity-coupling strength plane [152,214].

Let us finally mention a promising but challenging theoretical approach, which has the ambition to bridge the small-scale physics controlled by thermal nucleation of rupture to the large-scale organization of earthquakes and faults [155,205]. Partial success has been obtained with a remarkable prediction on the ("multifractal") dependence of the Omori law exponent of aftershocks on the magnitude of the mainshock, verified by careful empirical analyzes on earthquakes in California, Japan and worldwide [158].

Empirical Studies of Seismicity Inspired by Statistical Physics

> *"False facts are highly injurious to the progress of science, for they often endure long; but false views, if supported by some evidence, do little harm, for everyone takes a salutary pleasure in proving their falseness."*
>
> Charles Darwin, in The Origin of Man, Chap. 6.

Early Successes and Subsequent Challenges

A significant benefit of the statistical physics approach to seismology has been the introduction of novel techniques to analyze the available empirical data sets, with the goal of obtaining new insights into the spatio-temporal organization of seismicity and of revealing novel regularities and laws that may guide the theoretical analysis.

A prominent forerunner is the application of the concept of fractals introduced by Mandelbrot [137] and of the measures of fractal dimensions to describe complex sets of earthquake epicenters, hypocenters and fault patterns. The use of fractals has constituted an epistemologic breakthrough in many fields, and not only in seismology. Indeed, before Mandelbrot, when dealing with most complex systems, one used to say: "this is too complicated for a quantitative analysis" and only qualitative descriptions were offered. After Mandelbrot, one could hear: "this is a fractal, with a fractal dimension equal to xxx!" By providing a new geometrical way of thinking about complex systems associated with novel metrics, Mandelbrot and his fractals have extended considerably the reach of quantitative science to many complex systems in all fields.

However, while there have been some attempts to use fractal dimensions as guidelines to infer the underlying organization processes, as for instance in [195,196], most of the initial reports have lost their early appeal [18,19,183,219] since the complexity of seismicity and faulting is much too great to be captured by scaling laws embodying solely a simple scale invariance symmetry. Among others, multifractal and adapted wavelet tools are needed to quantify this complexity, see for instance [68,146,156,157,159]. It should also be noted that few studies of the fractal dimensions of seismicity address the significant issues of errors, biases and incomplete records in earthquake catalogs – a notable exception being [115].

Since the beginning of the 21st century, a renewal of interest and efforts have burgeoned as groups of statistical physicists, interested in earthquakes as a potential instance of self-organized criticality (SOC), have claimed "novel", "universal" and "robust" scaling laws from their analysis of the spatio-temporal organization of seismicity. The authors purport to have discovered universal and hitherto unknown features of earthquakes that give new insights into the dynamics of earthquakes and add to the evidence that earthquakes are self-organized critical. We now discuss a few of these recent studies to illustrate the existence of potential problems in the "statistical physics" approach. In a nutshell, we show that perhaps most of these "novel scaling laws" can be explained entirely by already known statistical seismicity laws. This claim has been defended by other experts of statistical seismology, the most vocal being perhaps Yan Kagan at UCLA [111].

The flurry of interest from physicists comes from their fascination with the self-similar properties exhibited by seismicity (e. g. the Gutenberg–Richter power law of earthquake seismic moments, the Omori–Utsu law of the decay of aftershock rates after large earthquakes, the fractal and multifractal space-time organization of earthquakes and faults, etc.), together with the development of novel concepts and techniques that may provide new insights. But, and this is our main criticism based on several detailed examples discussed below, many of the new approaches and results do not stand close scrutiny. This failure is rooted in two short-comings: (i) the lack of testing of new methods on synthetic catalogs generated by benchmark models which are based on well-known statistical laws of seismicity, and (ii) the failure to consider earthquake catalog bias, incompleteness and errors. The latter may cause catalog artifacts to appear as genuine characteristics of earthquakes. Testing the results on a variety of catalogs and considering the influence of various catalog errors can help minimize their influence. The former short-coming often leads to the following scenario: authors fail to realize that a simpler null hypothesis could not be rejected, namely that their "discovery" could actually be explained by just a combination of basic statistical laws known to seismologists for decades.

The well-established laws of statistical seismicity that authors should consider before claiming for novelty include the following:

1. The Gutenberg–Richter law for the distribution of earthquake magnitudes with a b-value close to 1 (corresponding to an exponent $\simeq 2/3$ for the probability density function of seismic moments),
2. The Omori–Utsu law for the decay of the rate of aftershocks following a mainshock,
3. The inverse Omori law for foreshocks,
4. The fact that aftershocks also trigger their own aftershocks and so on, and that aftershocks do not seem to exhibit any distinguishable physical properties,
5. The fact that the distribution of distances between mainshocks and aftershocks has a power law tail,
6. The fertility law (the fact that earthquakes of magnitude M trigger of the order of 10^{aM} aftershocks with $a\sim \leq b \simeq 1$,
7. The fractal distribution of faults which are concentration centers for earthquakes.

This above non-exhaustive list selects "laws" which are arguably non-redundant, in the sense that it is likely not possible to derive one of these laws from the others (a possible exception is the inverse Omori law for foreshock, which can be derived from the direct Omori law for aftershocks in the context of the ETAS model [78,81]). Some experts would argue that we should add to this list other claimed regularities, such as "Båth's law" (see e. g. [190] for a recent discussion emphasizing the importance of this law), that states that the differences in magnitudes between mainshocks and their largest aftershocks are approximately constant, independent of the magnitudes of mainshocks [20]. However, Helmstetter and Sornette [79] and Saichev and Sornette [173] have shown that Bath's law can be accurately recovered in ETAS-type models combining the first, second, fourth, and sixth laws stated above, with the assumption that any earthquake can trigger subsequent earthquakes.

Entropy Method for the Distribution of Time Intervals Between Mainshocks

Mega et al. [140] used the "diffusion entropy" method to argue for a power-law distribution of time intervals between a large earthquake (the mainshock of a seismic sequence or cluster) and the next one. Helmstetter and Sornette [80] showed that all the "new" discoveries reported by [140] (including the supposedly new scaling) can be explained solely by Omori's law for intra-cluster times, without correlation between clusters, thus debunking the claim for novelty.

Scaling of the PDF of Waiting Times

Bak et al. [14] analyzed the scaling of the probability density function of waiting times between successive earthquakes in southern California as a function of "box size" or small regions in which subsequent earthquakes are considered. They found an approximate collapse of the pdfs for different seismic moment thresholds S and box sizes L which suggested the following scaling ansatz for the waiting times T:

$$T^{\alpha} P_{S,L}(T) = f(TS^{-b}L^{d_f}) \,, \tag{3}$$

where $b = 1$ is the Gutenberg–Richter exponent, $d_f \simeq 1.2$ was claimed to be a spatial fractal dimension of seismicity (see [146] and [115] for more in-depth studies), $\alpha = 1$ was identified as the exponent in the Omori law and $f(\cdot)$ is a scaling function which was proposed to be roughly constant up to a constant ("kink") beyond which it quickly decays. The scaling (3) was claimed to be a unified law for earthquakes that revealed a novel feature in the spatio-temporal organization of seismicity in that the Gutenberg–Richter, the Omori law and the spatial distribution of earthquakes were unified into a single picture that made no distinction between fore-, main- and aftershocks. The scaling relations and critical exponents were claimed to be contained in the scaling ansatz. Corral [40,41,42,43] and others broadened the analysis to other regions of the world. Corral [41] proposed a slightly different scaling ansatz for a modified data analysis.

Early criticism came from Lindman et al. [132], who noted that synthetic data generated using a non-homogeneous Poisson process derived from Omori's law was able to reproduce some of the results of [14], indicating a rather trivial origin of the unified scaling law. Molchan [145] showed that, if at least two regions in the data set are independent, then, if a scaling relation were to hold exactly, this scaling function could only be exponential. All other functions could only result in approximate data collapses.

Proponents of the unified scaling law, e. g. [44], argued that indeed all regions were correlated, as expected in systems near a critical point so that the assumption of independence between different regions should not hold. But Molchan [145] also showed that a simple Poisson cluster model (Poissonian mainshocks that trigger Omori-type aftershock sequences) could reproduce the short and long time limits of the observed statistics, indicating that the Omori law, the Gutenberg–Richter relationship and simple clustering were the sole ingredients necessary for the observed short and long time limit, and no spatial correlation was needed.

Saichev and Sornette [176,177] extended Molchan's arguments to show that the approximate data collapse of the waiting times could be explained completely by the Epidemic-Type Aftershock Sequence (ETAS) model of [149]. This provided further evidence that the apparent data collapse was only approximate. Remarkably, the theoretical predictions of the ETAS model seem to fit the observed data better than the phenomenological scaling function proposed by [41] to fit the data. Saichev and Sornette [176,177] thus showed that a benchmark model of seismicity was able to reproduce the apparent unified scaling law and that therefore the distribution of interevent times did not reveal new information beyond what was already known via statistical laws: The combination of the Gutenberg–Richter law, the Omori law, and the concept of clustering suffice to explain the apparent "universal" scaling of the waiting times.

Sornette et al. [215] developed an efficient numerical scheme to solve accurately the set of nonlinear integral equations derived previously in [177] and found a dramatic lack of power for the distribution of inter-event times to distinguish between quite different sets of parameters, casting doubt on the usefulness of this statistics for the specific purpose of identifying the clustering parameter (e. g. [72]).

Scaling of the PDF of Distances Between Subsequent Earthquakes

Davidson and Paczuski [49] claimed evidence contradicting the theory of aftershock zone scaling in favor of scale-free statistics. Aftershock zone scaling refers to the scaling of the mainshock rupture length, along which most aftershocks occur, with the mainshock magnitude [112]. Davidson and Paczuski [49] suggested that the probability density function of spatial distances between successive earthquakes obeys finite size scaling with a novel dynamical scaling exponent, suggesting that the mainshock rupture length scale has no impact on the spatial distribution of aftershocks and that earthquakes are self-organized critical.

Werner and Sonette [230] debunked this claim by showing that (i) the purported power law scaling function is not universal as it breaks down in other regions of the world; (ii) the results obtained by [49] for southern California depend crucially on a single earthquake (the June 28, 1992, M7.3 Landers earthquake): without Landers and its aftershocks, the power law disappears; (iii) a model of clustered seismicity, with aftershock zone scaling explicitly built in, is able reproduce the apparent power law, indicating that an apparent lack of scales in the data does not necessarily contradict aftershock zone scaling and the existence of scales associated with mainshock rupture length scales.

The Network Approach

The recent boom in the statistical mechanics of network analysis has recently extended to applications well beyond physics (for reviews, see [6,16,57,148]). Earthquake seismology is no exception [1,2,3,4,8,9,10,160,161]. The resulting impact has been limited so far for several reasons.

A major concern is the assumption that earthquake catalogs as downloaded from the web are data sets fit for immediate analysis. References [82,113,229] and [231] present modern and complementary assessments of the many issues of incompleteness spoiling even the best catalogs. In particular, we should stress that magnitude determinations are surprisingly inaccurate, leading to large errors in seismic rate estimates [231]. Furthermore, there is no such thing as a complete catalog above a so-called magnitude of completeness, due to the fact that a non-negligible fraction of earthquakes are missed in the aftermath of previous earthquakes [82,113]. One should be concerned that analyses in terms of network metrics could be particularly sensitive to these defects. Nevertheless, Abe and Suzuki [1,2,3,4] applied metrics of network analysis to "raw" catalogs which included events well below the estimated magnitude of completeness. As a result of neglecting to use a (reasonably) homogeneous and trustworthy data set, the results of their analysis may be severely biased, because the reliability of the inferred network structure is probably more sensitive than other metrics to the correct spatio-temporal ordering of the earthquake catalog. No serious study has yet been performed to quantify the usually serious impact of quality issues on the metrics used in network analysis. As a consequence, it is also not clear how to interpret the "success" of [160,161] in reproducing the "features" of Abe and Suzuki's analysis on the synthetic seismicity generated by a spring-block model.

In addition, at best limited attempts have been made to interpret the results of the new network metrics using well-known, established facts in seismology. Many of the claimed novel features are probably very well understood – they are mostly related to scale-invariance and clustering of seismicity, facts documented for decades. The authors should always strive to show that the new metrics that they propose give results that cannot be explained by the standard laws in statistical seismology. Toward this end, there are well-defined benchmark models that incorporate these laws and that can generate synthetic catalogs on which the new metrics can be tested and compared.

A few exceptions are worth mentioning. Motivated by the long-standing and unresolved debate over "aftershock" identification, Baiesi and Paczuski [9,10] and Baiesi [8] provided a new metric for the correlations between earthquakes based on the space-time-magnitude nearest-neighbor distance between earthquakes. The authors compared their results with known statistical laws in seismology and with the predictions of the ETAS model, actually confirming both. While no new law has been unearthed here, such efforts are valuable to validate known laws and continue to test the possible limits. Zaliapin et al. [238] extended their study and investigated the theoretical properties of the metric and its ability to decluster catalogs (i. e., separate mainshocks from aftershocks). They concluded that aftershocks defined from this metric seem to be different from the rest of earthquakes. It will be interesting to see head-to-head comparisons with current state-of-the-art probabilistic declustering techniques that are based on empirical statistical laws and likelihood estimation [105, 240,241].

Future Directions

The study of the statistical physics of earthquakes remains wide-open with many significant discoveries to be made. The promise of a holistic approach – one that emphasizes the interactions between earthquakes and faults – is to be able to neglect some of the exceedingly complicated micro-physics when attempting to understand the large scale patterns of seismicity. The marriage between this conceptual approach, based on the successes of statistical physics, and seismology thus remains a highly important domain of research. In particular, statistical seismology needs to evolve into a genuine physically-based statistical physics of earthquakes.

The question of renormalizability of models of earthquake occurrence and the role of small earthquakes in the organization of seismicity is likely to remain an important topic. It connects with the problem of foreshocks and the predictability of large events from small ones and therefore has real and immediate practical applications as well as physical implications.

More detailed and rigorous empirical studies of the frequency-size statistics of earthquake seismic moments and how they relate to seismo-tectonic conditions are needed in order to help settle the controversy over the power-law versus the characteristic event regime, and the role of regime-switching and universality.

Spatially extended, dynamically evolving fault networks and their role in the generation of earthquakes are mostly ignored in the statistical physics approach to seismicity. Akin to the filaments in turbulence, these may provide key insights into the spatio-temporal organization of earthquakes. Novel methods combining information from seismology to faulting will be required (e. g., [71,159,195, 196,197]) to build a real understanding of the self-organization of the chicken-and-egg structures that earthquakes-faults constitute. Furthermore, a true physical approach requires understanding the spatio-temporal evolution of stresses, their role in earthquake nucleation via thermally activated processes, in the rupture propagation and in the physics of arrest, both involved in the generation of complex stress fields.

The important debate regarding statistical physics approaches to seismicity would benefit significantly from two points. Firstly, earthquake catalogs contain data uncertainties, biases and subtle incompleteness issues. Investigating their influence on the results of data analyses inspired by statistical physics increases the relevance of the results. Secondly, the authors should make links with the literature on statistical seismology which deals with similar questions. It is their task to show that the new metrics that they propose give results that cannot be explained by the standard laws in statistical seismology. For this, there are well-defined benchmark models that incorporate these laws and that can generate synthetic catalogs on which the new metrics can be tested.

Bibliography

1. Abe S, Suzuki N (2004) Scale-free network of earthquakes. Europhys Lett 65:581–586. doi:10.1209/epl/i2003-10108-1
2. Abe S, Suzuki N (2004) Small-world structure of earthquake network. Physica A: Stat Mech Appl 337:357–362. doi:10.1016/j.physa.2004.01.059
3. Abe S, Suzuki N (2005) Scale-invariant statistics of period in directed earthquake network. Eur Phys J B 44:115–117. doi:10.1140/epjb/e2005-00106-7
4. Abe S, Suzuki N (2006) Complex earthquake networks: Hierarchical organization and assortative mixing. Phys Rev E 74(2):026, 113–+. doi:10.1103/PhysRevE.74.026113

5. Aki K (1995) Earthquake prediction, societal implications. Rev Geophys 33:243–248
6. Albert R, Barabási AL (2002) Statistical mechanics of complex networks, Rev Mod Phys 74(1):47–97. doi:10.1103/RevModPhys.74.47
7. Allègre CJ, Le Mouel JL, Provost A (1982) Scaling rules in rock fracture and possible implications for earthquake prediction. Nature 297:47–49. doi:10.1038/297047a0
8. Baiesi M (2006) Scaling and precursor motifs in earthquake networks. Physica A: Stat Mech Appl 359:775–783. doi:10.1016/j.physa.2005.05.094
9. Baiesi M, Paczuski M (2004) Scale-free networks of earthquakes and aftershocks. Phys Rev E 69(6):066, 106. doi:10.1103/PhysRevE.69.066106
10. Baiesi M, Paczuski M (2005) Complex networks of earthquakes and aftershocks. Nonlin Proc Geophys 12:1–11
11. Bak P (1996) How Nature Works: The Science of Self-Organized Criticality. Springer, New York, p 212
12. Bak P, Tang C (1989) Earthquakes as a self-organized critical phenomena. J Geophys Res 94(B11):15635–15637
13. Bak P, Tang C, Wiesenfeld K (1987) Self-organized criticality: An explanation of the 1/*f* noise. Phys Rev Lett 59(4):381–384. doi:10.1103/PhysRevLett.59.381
14. Bak P, Christensen K, Danon L, Scanlon T (2002) Unified scaling law for earthquakes. Phys Rev Lett 88(17):178,501. doi:10.1103/PhysRevLett.88.178501
15. Bakun, WH, Aagaard B, Dost B, Ellsworth WL, Hardebeck JL, Harris RA, Ji C, Johnston MJS, Langbein J, Lienkaemper JJ, Michael AJ, Murray JR, Nadeau RM, Reasenberg PA, Reichle MS, Roeloffs EA, Shakal A, Simpson RW, Waldhauser F (2005) Implications for prediction and hazard assessment from the 2004 Parkfield earthquake. Nature 437:969–974. doi:10.1038/nature04067
16. Barabási AL, Albert R (1999) Emergence of Scaling in Random Networks. Science 286(5439):509–512. doi:10.1126/science.286.5439.509
17. Barabási AL, Albert R, Jeong H (1999) Mean-field theory fore scale-free random networks. Physica A 272:173–187. doi:10.1016/S0378-4371(99)00291-5
18. Barton CC, La Pointe PR (eds) (1995) Fractals in the Earth Sciences. Plenum Press, New York, London
19. Barton CC, La Pointe PR (eds) (1995) Fractals in petroleum geology and earth processes. Plenum Press, New York, London
20. Båth M (1965) Lateral inhomogeneities in the upper mantle. Tectonophysics 2:483–514
21. Ben-Zion Y, Rice JR (1993) Earthquake failure sequences along a cellular fault zone in a 3-dimensional elastic solid containing asperity and nonasperity regions. J Geophys Res 93:14109–14131
22. Ben-Zion Y, Rice JR (1995) Slip patterns and earthquake populations along different classes of faults in elastic solids. J Geophys Res 100:12959–12983
23. Ben-Zion Y, Rice JR (1997) Dynamic simulations of slip on a smooth fault in an elastic solid. J Geophys Res 102:17771–17784
24. Ben-Zion Y, Dahmen K, Lyakhovsky V, Ertas D, Agnon A (1999) Self-driven mode switching of earthquake activity on a fault system. Earth Planet Sci Lett 172:11–21
25. Bird P, Kagan YY (2004) Plate-tectonic analysis of shallow seismicity: Apparent boundary width, beta, corner magnitude, coupled lithosphere thickness, and coupling in seven tectonic settings. Bull Seismol Soc Am 94(6):2380–2399
26. Bowman DD, King GCP (2001) Stress transfer and seismicity changes before large earthquakes. C Royal Acad Sci Paris, Sci Terre Planetes 333:591–599
27. Bowman DD, King GCP (2001) Accelerating seismicity and stress accumulation before large earthquakes. Geophys Res Lett 28:4039–4042
28. Bowman DD, Oullion G, Sammis CG, Sornette A, Sornette D (1998) An observational test of the critical earthquake concept. J Geophys Res 103:24359–24372
29. Brehm DJ, Braile LW (1998) Intermediate-term earthquake prediction using precursory events in the New Madrid Seismic Zone. Bull Seismol Am Soc 88(2):564–580
30. Bufe CG, Varnes DJ (1993) Predictive modeling of the seismic cycle of the greater San Francisco Bay region. J Geophys Res 98:9871–9883
31. Burridge R, Knopoff L (1964) Body force equivalents for seismic dislocation. Seism Soc Am Bull 54:1875–1888
32. Cardy JL (1996) Scaling and Renormalization in Statistical Physics. Cambridge University Press, Cambridge
33. Carlson JM, Langer JS (1989) Properties of earthquakes generated by fault dynamics. Phys Rev Lett 62:2632–2635
34. Carlson JM, Langer JS, Shaw BE (1994) Dynamics of earthquake faults. Rev Mod Phys 66:657–670
35. Chelidze TL (1982) Percolation and fracture. Phys Earth Planet Interiors 28:93–101
36. Christensen K, Farid N, Pruessner G, Stapleton M (2008) On the finite-size scaling of probability density functions. Eur Phys B 62:331–336
37. Clauset A, Shalizi CR, Newman MEJ (2007) Power-law distributions in empirical data. E-print arXiv:0706.1062
38. Cochard A, Madariaga R (1994) Dynamic faulting under rate-dependent friction. Pure Appl Geophys 142:419–445
39. Cochard A, Madariaga R (1996) Complexity of seismicity due to highly rate-dependent friction. J Geophys Res 101:25321–25336
40. Corral A (2003) Local distributions and rate fluctuations in a unified scaling law for earthquakes. Phys Rev E 68(3):035, 102. doi:10.1103/PhysRevE.68.035102
41. Corral A (2004) Universal local versus unified global scaling laws in the statistics of seismicity. Physica A 340:590–597
42. Corral A (2004) Long-term clustering, scaling, and universality in the temporal occurrence of earthquakes. Phys Rev Lett 92:108, 501
43. Corral A (2005) Mixing of rescaled data and bayesian inference for earthquake recurrence times. Nonlin Proc Geophys 12:89–100
44. Corral A (2005) Renormalization-group transformations and correlations of seismicity. Phys Rev Lett 95:028, 501
45. Corral A (2006) Universal earthquake-occurrence jumps, correlations with time, and anomalous diffusion. Phys Rev Lett 97:178, 501
46. Corral A, Christensen K (2006) Comment on "earthquakes descaled: On waiting time distributions and scaling laws". Phys Rev Lett 96:109, 801
47. Dahmen K, Ertaş D, Ben-Zion Y (1998) Gutenberg–Richter and characteristic earthquake behavior in simple mean-field models of heterogeneous faults. Phys Rev E 58:1494–1501. doi:10.1103/PhysRevE.58.1494

48. Davidsen J, Goltz C (2004) Are seismic waiting time distributions universal? Geophys Res Lett 31:L21612. doi:10.1029/2004GL020892
49. Davidsen J, Paczuski M (2005) Analysis of the spatial distribution between successive earthquakes. Phys Rev Lett 94:048, 501. doi:10.1103/PhysRevLett.94.048501
50. Davidsen J, Grassberger P, Paczuski M (2006) Earthquake recurrence as a record breaking process. Geophys Res Lett 33:L11304. doi:10.1029/2006GL026122
51. Dhar D (1990) Self-organized critical state of sandpile automaton models. Phys Rev Lett 64:1613–1616
52. Dhar D (1999) The Abelian sandpile and related models. Physica A 263:4–25
53. Dieterich JH (1987) Nucleation and triggering of earthquake slip; effect of periodic stresses. Tectonophysics 144:127–139
54. Dieterich JH (1992) Earthquake nucleation on faults with rate-dependent and state-dependent strength. Tectonophysics 211:115–134
55. Dieterich J (1994) A constitutive law for rate of earthquake production and its application to earthquake clustering. J Geophys Res 99:2601–2618
56. Dieterich J, Kilgore BD (1994) Direct observation of frictional constacts- New insight for state-dependent properties. Pure Appl Geophys 143:283–302
57. Dorogevtsev SN, Mendes JFF (2003) Evolution of Networks: From Biological Nets to the Internet and WWW. Oxford University Press, New York
58. Eckman JP (1981) Roads to Turbulence in Dissipative Dynamical Systems. Rev Mod Phys 53:643–654
59. Ellsworth WL, Lindh AG, Prescott WH, Herd DJ (1981) The 1906 San Francisco Earthquake and the seismic cycle. Am Geophys Union Maurice Ewing Monogr 4:126–140
60. Felzer KR, Becker TW, Abercrombie RE, Ekstrom G, Rice JR (2002) Triggering of the 1999 Mw 7.1 Hector Mine earthquake by aftershocks of the 1992 Mw 7.3 Landers earthquake. J Geophys Res 107(B09):2190
61. Fisher DS, Dahmen K, Ramanathan S, Ben-Zion Y (1997) Statistics of Earthquakes in Simple Models of Heterogeneous Faults. Phys Rev Lett 78:4885–4888. doi:10.1103/PhysRevLett.78.4885
62. Freund F, Sornette D (2007) Electro-Magnetic Earthquake Bursts and Critical Rupture of Peroxy Bond Networks in Rocks. Tectonophysics 431:33–47
63. Frisch U (1995) Turbulence. The legacy of A.N. Kolmogorov. Cambridge University Press, Cambridge
64. Gabrielov A, Keilis-Borok V, Jackson DD (1996) Geometric Incompatibility in a Fault System. Proc Nat Acad Sci 93:3838–3842
65. Gabrielov A, Keilis-Borok V, Zaliapin I, Newman W (2000) Critical transitions in colliding cascades. Phys Rev E 62:237–249
66. Gabrielov A, Zaliapin I, Newman W, Keilis-Borok V, (2000) Colliding cascades model for earthquake prediction. Geophys J Int 143:427–437
67. Gallagher R, Appenzeller T (1999) Beyond Reductionism. Science 284(5411):79
68. Geilikman MB, Pisarenko VF, Golubeva TV (1990) Multifractal Patterns of Seismicity. Earth Planet Sci Lett 99:127–138
69. Gelfand IM, Guberman SA, Keilis-Borok VI, Knopoff L, Press F, Ranzman EY, Rotwain IM, Sadovsky AM (1976) Pattern recognition applied to earthquake epicenters in California. Phys Earth Planet Interiors 11:227–283
70. Geller RJ, Jackson DD, Kagan YY, Mulargia F (1997) Earthquakes cannot be predicted. Science 275:1616–1617
71. Gorshkov A, Kossobokov V, Soloviev A (2003) Recognition of earthquake-prone areas. In: Keilis-Borok V, Soloviev A (eds) Nonlinear Dynamics of the Lithosphere and Earthquake Prediction. Springer, Heidelberg, pp 239–310 [122]
72. Hainzl S, Scherbaum F, Beauval C (2006) Estimating Background Activity Based on Interevent-Time Distribution. Bull Seismol Soc Am 96(1):313–320. doi:10.1785/0120050053
73. Hanks TC (1992) Small earthquakes, tectonic forces. Science 256:1430–1432
74. Hardebeck JL, Felzer KR, Michael AJ (2008) Improved tests reveal that the accelerating moment release hypothesis is statistically insignificant. J Geophys Res 113:B08310. doi:10.1029/2007JB005410
75. Harris RA, Arrowsmith JR (2006) Introduction to the Special Issue on the 2004 Parkfield Earthquake and the Parkfield Earthquake Prediction Experiment. Bull Seismol Soc Am 96(4B):S1–10. doi:10.1785/0120050831
76. Helmstetter A (2003) Is earthquake triggering driven by small earthquakes? Phys Rev Lett 91(5):058, 501. doi:10.1103/PhysRevLett.91.058501
77. Helmstetter A, Sornette D (2002) Subcritical and supercritical regimes in epidemic models of earthquake aftershocks. J Geophys Res 107(B10):2237. doi:10.1029/2001JB001580
78. Helmstetter A, Sornette D (2003) Foreshocks explained by cascades of triggered seismicity. J Geophys Res (Solid Earth) 108(B10):2457 doi:10.1029/2003JB00240901
79. Helmstetter A, Sornette D (2003) Bath's law Derived from the Gutenberg–Richter law and from Aftershock Properties. Geophys Res Lett 30:2069. doi:10.1029/2003GL018186
80. Helmstetter A, Sornette D (2004) Comment on "Power-Law Time Distribution of Large Earthquakes". Phys Rev Lett 92:129801 (Reply is Phys Rev Lett 92:129802 (2004))
81. Helmstetter A, Sornette D, Grasso J-R (2003) Mainshocks are Aftershocks of Conditional Foreshocks: How do foreshock statistical properties emerge from aftershock laws. J Geophys Res 108(B10):2046. doi:10.1029/2002JB001991
82. Helmstetter A, Kagan YY, Jackson DD (2005) Importance of small earthquakes for stress transfers and earthquake triggering. J Geophys Res 110:B05508. doi:10.1029/2004JB003286
83. Helmstetter A, Kagan Y, Jackson D (2006) Comparison of short-term and long-term earthquake forecast models for Southern California. Bull Seism Soc Am 96:90–106
84. Hergarten S (2002) Self-Organized Criticality in Earth Systems. Springer, Berlin
85. Hinrichsen H (2000) Non-equilibrium critical phenomena and phase transitions into absorbing states. Adv Phys 49:815–958(144)
86. Holliday JR, Rundle JB, Tiampo KF, Klein W, Donnellan A (2006) Systematic procedural and sensitivity analysis of the Pattern Informatics method for forecasting large ($M > 5$) earthquake events in Southern California. Pure Appl Geophys 163(11–12):2433–2454
87. Huang J, Turcotte DL (1990) Evidence for chaotic fault interactions in the seismicity of the San Andreas fault and Nankai trough. Nature 348:234–236
88. Huang J, Turcotte DL (1990) Are earthquakes an example of deterministic chaos? Geophys Rev Lett 17:223–226

89. Huang Y, Saleur H, Sammis CG, Sornette D (1998) Precursors, aftershocks, criticality and self-organized criticality. Europhys Lett 41:43–48
90. Ide K, Sornette D (2002) Oscillatory Finite-Time Singularities in Finance, Population and Rupture. Physica A307(1–2):63–106
91. Jackson DD, Kagan YY (2006) The 2004 Parkfield Earthquake, the 1985 Prediction, and Characteristic Earthquakes: Lessons for the Future. Bull Seismol Soc Am 96(4B):S397–409. doi:10.1785/0120050821
92. Jaumé SC, Sykes LR (1999) Evolving Towards a Critical Point: A Review of Accelerating Seismic Moment/Energy Release Prior to Large and Great Earthquakes. Pure Appl Geophys 155:279–305
93. Jensen HJ (1998) Self-Organized Criticality: Emergent Complex Behavior in Physical and Biological Systems. Cambridge University Press, Cambridge
94. Johansen A, Sornette D, Wakita G, Tsunogai U, Newman WI, Saleur H (1996) Discrete scaling in earthquake precursory phenomena: evidence in the Kobe earthquake, Japan J Phys I France 6:1391–1402
95. Johansen A, Saleur H, Sornette D (2000) New Evidence of Earthquake Precursory Phenomena in the 17 Jan. 1995 Kobe Earthquake, Japan. Eur Phys J B 15:551–555
96. Jones LM (1994) Foreshocks, aftershocks, and earthquake probabilities: accounting for the Landers earthquake. Bull Seismol Soc Am 84:892–899
97. Jordan TH (2006) Earthquake Predictability, Brick by Brick. Seismol Res Lett 77(1):3–6
98. Kadanoff LP, Nagel SR, Wu L, Zhou S-M (1989) Scaling and universality in avalanches. Phys Rev A 39(12):6524–6537. doi:10.1103/PhysRevA.39.6524
99. Kagan YY (1981), Spatial distribution of earthquakes: The three-point moment function. Geophys J R Astron Soc 67:697–717
100. Kagan YY (1981) Spatial distribution of earthquakes: The four-point moment function. Geophys J Roy Astron Soc 67:719–733
101. Kagan YY (1987) Point sources of elastic deformation: Elementary sources, static displacements. Geophys J R Astron Soc 90:1–34
102. Kagan YY (1987) Point sources of elastic deformation: Elementary sources, dynamic displacements. Geophys J R Astron Soc 91:891–912
103. Kagan YY (1988) Multipole expansions of extended sources of elastic deformation. Geophys J R Astron Soc 93:101–114
104. Kagan YY (1989) Earthquakes and fractals. Ann Rev Mater Sci: Fractal Phenom Disordered Syst 19:520–522
105. Kagan YY (1991) Likelihood analysis of earthquake catalogs. Geophys J Int 106:135–148
106. Kagan YY (1992) Seismicity: Turbulence of solids. Nonlinear Sci Today 2:1–13
107. Kagan YY (1992) On the geometry of an earthquake fault system. Phys Earth Planet Interiors 71:15–35
108. Kagan YY (1993) Statistics of characteristic earthquakes. Bull Seismol Soc Am 83(1):7–24
109. Kagan YY (1994) Observational evidence for earthquakes as a nonlinear dynamic process. Physica D 77:160–192
110. Kagan YY (1994) Comment on "The Gutenberg–Richter or char-acteristic earthquake distribution, which is it?" by Wesnousky. Bull Seismol Soc Am 86:274–285
111. Kagan YY (1999) Is earthquake seismology a hard, quantitative science? Pure Appl Geophys 155:33–258
112. Kagan YY (2002) Aftershock Zone Scaling. Bull Seismol Soc Am 92(2):641–655. doi:10.1785/0120010172
113. Kagan YY (2003) Accuracy of modern global earthquake catalogs. Phys Earth Planet Interiors 135:173–209
114. Kagan YY (2006) Why does theoretical physics fail to explain and predict earthquake occurrence? In: Bhattacharyya P, Chakrabarti BK (eds) Modelling Critical and Catastrophic Phenomena in Geoscience: A Statistical Physics Approach. Lecture Notes in Physics, vol 705. Springer, Berlin, pp 303–359
115. Kagan YY (2007) Earthquake spatial distribution: the correlation dimension. Geophys J Int 168:1175–1194. doi:10.1111/j.1365-246X.2006.03251.x
116. Kagan YY, Knopoff L (1980) Spatial distribution of earthquakes: The two-point correlation function. Geophys J R Astron Soc 62:303–320
117. Kagan YY, Knopoff L (1981) Stochastic synthesis of earthquake catalogs. J Geophys Res 86(B4):2853–2862
118. Kagan YY, Knopoff L (1985) The first-order statistical moment of the seismic moment tensor. Geophys J R Astron Soc 81:429–444
119. Kagan YY, Knopoff L (1985) The two-point correlation function of the seismic moment tensor. Geophys J R Astron Soc 83:637–656
120. Keilis-Borok VI (ed) (1990) Intermediate-term earthquake prediction: models, algorithms, worldwide tests. Phys Earth Planet Interiors 61(1–2)
121. Keilis-Borok VI, Malinovskaya LN (1964) One regularity in the occurrence of strong earthquakes. J Geophys Res B 69:3019–3024
122. Keilis-Borok V, Soloviev A (2003) Nonlinear Dynamics of the Lithosphere and Earthquake Prediction. Springer, Heidelberg
123. Keilis-Borok VI, Knopoff L, Rotwain IM, Allen CR (1988) Intermediate-term prediction of occurrence times of strong earthquakes. Nature 335:690–694
124. King GCP, Bowman DD (2003) The evolution of regional seismicity between large earthquakes. J Geophys Res 108(B2):2096. doi:10.1029/2001JB000783
125. Klein W, Rundle JB, Ferguson CD (1997) Scaling and nucleation in models of earthquake faults. Phys Rev Lett 78:3793–3796
126. Knopoff L (1996) The organization of seismicity on fault networks. Proc Nat Acad Sci USA 93:3830–3837
127. Landau LD, Lifshitz EM (1980) Statistical Physics Course on Theoretical Physics, vol 5, 3rd edn. Butterworth-Heinemann, Oxford
128. Langer JS, Carlson JM, Myers CR, Shaw BE (1996) Slip complexity in dynamical models of earthquake faults. Proc Nat Acad Sci USA 93:3825–3829
129. Lee MW, Sornette D, Knopoff L (1999) Persistence and Quiescence of Seismicity on Fault Systems. Phys Rev Lett 83(N20):4219–4222
130. Levin SZ, Sammis CG, Bowman DD (2006) An observational test of the stress accumulation model based on seismicity preceding the 1992 Landers, CA earthquake. Tectonophysics 413:39–52
131. Lindh AG (1990) The seismic cycle pursued. Nature 348:580–581

132. Lindman M, Jonsdottir K, Roberts R, Lund B, Bdvarsson R (2005) Earthquakes descaled: On waiting time distributions and scaling laws. Phys Rev Lett 94:108, 501
133. Lindman M, Jonsdottir K, Roberts R, Lund B, Bdvarsson R (2006) Reply to comment by A. Corral and K. Christensen. Phys Rev Lett 96:109, 802
134. Livina VN, Havlin S, Bunde A (2006) Memory in the occurrence of earthquakes. Phys Rev Lett 95:208, 501
135. Luebeck S (2004) Universal scaling behavior of non-equilbrium phase transitions. Int J Mod Phys B 18:3977
136. Manna S (1991) Critical exponents of the sandpile models in two dimensions. Physica A179(2):249–268
137. Mandelbrot BB (1982) The Fractal Geometry of Nature. W.H. Freeman, San Francisco
138. Marsan D (2005) The role of small earthquakes in redistributing crustal elastic stress. Geophys J Int 163(1):141–151. doi:10.1111/j.1365-246X.2005.02700.x
139. May RM (1976) Simple mathematical models with very complicated dynamics. Nature 261:459–467
140. Mega MS, Allegrini P, Grigolini P, Latora V, Palatella L, Rapisarda A, Vinciguerra S (2003) Power law time distributions of large earthquakes. Phys Rev Lett 90:18850
141. Michael AJ, Jones LM (1998) Seismicity alert probabilities at Parkfield, California, revisited. Bull Seismol Soc Am 88(1):117–130
142. Miltenberger P, Sornette D, Vanneste C (1993) Fault self-organization as optimal random paths selected by critical spatiotemporal dynamics of earthquakes. Phys Rev Lett 71:3604–3607. doi:10.1103/PhysRevLett.71.3604
143. Mitzenmacher M (2004) A Brief History of Generative Models for Power Law and Lognormal Distributions. Internet Math 1(2):226–251
144. Mogi K (1969) Some features of recent seismic activity in and near Japan 2: activity before and after great earthquakes. Bull Eq Res Inst Tokyo Univ 47:395–417
145. Molchan G (2005) Interevent time distribution in seismicity: A theoretical approach. Pure Appl Geophys 162:1135–1150. doi:10.1007/s00024-004-2664-5
146. Molchan G, Kronrod T (2005) On the spatial scaling of seismicity rate. Geophys J Int 162(3):899–909. doi:10.1111/j.1365-246X.2005.02693.x
147. Nature Debates (1999) Nature debates: Is the reliable prediction of individual earthquakes a realistic scientific goal? available from http://www.nature.com/nature/debates/earthquake/equake_frameset.html
148. Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256. doi:10.1137/S003614450342480
149. Ogata Y (1988) Statistical models for earthquake occurrence and residual analysis for point processes. J Am Stat Assoc 83:9–27
150. Ogata Y (1998) Space-time point-process models for earthquake occurrences. Ann Inst Stat Math 5(2):379–402
151. Olami Z, Feder HJS, Christensen K (1992) Self-organized criticality in a continuous, nonconservative cellular automaton modeling earthquakes. Phys Rev Lett 68(8):1244–1247
152. Osorio I, Frei MG, Sornette D, Milton J, Lai Y-C (2007) Seizures and earthquakes: Universality and scaling of critical far from equilibrium systems. submitted to Phys Rev Lett. http://arxiv.org/abs/0712.3929
153. Ouillon G, Sornette D (2000) The critical earthquake concept applied to mine rockbursts with time-to-failure analysis. Geophys J Int 143:454–468
154. Ouillon G, Sornette D (2004) Search for Direct Stress Correlation Signatures of the Critical Earthquake Model. Geophys J Int 157:1233–1246
155. Ouillon G, Sornette D (2005) Magnitude-Dependent Omori Law: Theory and Empirical Study. J Geophys Res 110:B04306. doi:10.1029/2004JB003311
156. Ouillon G, Sornette D, Castaing C (1995) Organization of joints and faults from 1 cm to 100 km scales revealed by Optimized Anisotropic Wavelet Coefficient Method and Multifractal analysis. Nonlinear Process Geophys 2:158–177
157. Ouillon G, Castaing C, Sornette D (1996) Hierarchical scaling of faulting. J Geophys Res 101(B3):5477–5487
158. Ouillon G, Ribeiro E, Sornette D (2007) Multifractal Omori Law for Earthquake Triggering: New Tests on the California, Japan and Worldwide Catalogs. submitted to Geophys J Int. http://arxiv.org/abs/physics/0609179
159. Ouillon G, Ducorbier C, Sornette D (2008) Automatic reconstruction of fault networks from seismicity catalogs: Three-dimensional optimal anisotropic dynamic clustering. J Geophys Res 113:B01306. doi:10.1029/2007JB005032
160. Peixoto TP, Prado CP (2004) Distribution of epicenters in the Olami–Feder–Christensen model. Phys Rev E 69(2):025101. doi:10.1103/PhysRevE.69.025101
161. Peixoto TP, Prado CPC (2006) Network of epicenters of the Olami–Feder–Christensen model of earthquakes. Phys Rev E 74(1):016, 126 doi:10.1103/PhysRevE.74.016126
162. Peters O, Neelin JD (2006) Critical phenomena in atmospheric precipitation. Nature Phys 2:393–396. doi:10.1038/nphys314
163. Pruessner G (2004) Studies in self-organized criticality, Ph D thesis, Imperial College London, available from http://www.ma.imperial.ac.uk/%7Epruess/publications/thesis_final/
164. Raleigh CB, Sieh K, Sykes LR, Anderson DL (1982) Forecasting Southern California Earthquakes. Science 217:1097–1104
165. Reynolds PJ, Klein W, Stanley HE (1977) Renormalization Group for Site and Bond Percolation. J Phys C 10:L167–L172
166. Rhoades DA, Evison FF (2004) Long-range earthquake forecasting with every earthquake a precursor according to scale. Pure Appl Geophys 161:47–72
167. Rhoades DA, Evison FF (2005) Test of the EEPAS forecasting model on the Japan earthquake catalogue. Pure Appl Geophys 162:1271–1290
168. Rice JR (1993) Spatio-temporal complexity of slip on a fault. J Geophys Res 98:9885–9907
169. Rundle JB, Klein W (1993) Scaling and critical phenomena in a cellular automaton slider block model for earthquakes. J Stat Phys 72:405–412
170. Rundle JB, Klein W (1995) New ideas about the physics of earthquakes. Rev Geophys 33:283–286
171. Rundle PB, Rundle JB, Tiampo KF, Sa Martins JS, McGinnis S, Klein W (2001) Nonlinear network dynamics on earthquake fault systems. Phys Rev Lett 87(14):148, 501. doi:10.1103/PhysRevLett.87.148501
172. Rundle JB, Turcotte DL, Shcherbakov R, Klein W, Sammis C (2003) Statistical physics approach to understanding the multiscale dynamics of earthquake fault systems. Rev Geophys 41(4):1019

173. Saichev A, Sornette D (2005) Distribution of the Largest Aftershocks in Branching Models of Triggered Seismicity: Theory of the Universal Bath's law. Phys Rev E 71:056127
174. Saichev A, Sornette D (2005) Vere-Jones' self-similar branching model. Phys Rev E 72:056, 122
175. Saichev A, Sornette D (2006) Renormalization of branching models of triggered seismicity from total to observable seismicity. Eur Phys J B 51:443–459
176. Saichev A, Sornette D (2006) "Universal" distribution of interearthquake times explained. Phys Rev Lett 97:078, 501
177. Saichev A, Sornette D (2007). Theory of earthquake recurrence times. J Geophys Res 112:B04313. doi:10.1029/2006JB004536
178. Saleur H, Sammis CG, Sornette D (1996) Renormalization group theory of earthquakes. Nonlinear Process Geophys 3:102–109
179. Saleur H, Sammis CG, Sornette D (1996) Discrete scale invariance, complex fractal dimensions and log-periodic corrections in earthquakes. J Geophys Res 101:17661–17677
180. Sammis SG, Sornette D (2002) Positive Feedback, Memory and the Predictability of Earthquakes. Proc Nat Acad Sci USA V99:SUPP1:2501–2508
181. Scholz CH (1991) Earthquakes and faulting: Self-organized critical phenomena with a characteristic dimension. In: Riste T, Sherrington D (eds) Spontaneous Formation of Space Time Structure and Criticality. Kluwer, Norwell, pp 41–56
182. Scholz CH (2002) The Mechanics of Earthquakes and Faulting, 2nd edn, Cambridge University Press, Cambridge
183. Scholz CH, Mandelbrot BB (eds) (1989) Fractals in Geophysics. Birkhäuser, Basel
184. Schorlemmer D, Wiemer S, Wyss M (2005) Variations in earthquake-size distribution across different stress regimes. Nature 437:539–542. doi:10.1038/nature04094
185. Schwartz DP, Coppersmith KJ (1984) Fault behavior and characteristic earthquakes: examples from the Wasatch and San Andreas Fault Zones. J Geophys Res 89:5681–5698
186. Shaw BE (1993) Generalized Omori law for aftershocks and foreshocks from a simple dynamics. Geophys Res Lett 20:907–910
187. Shaw BE (1994) Complexity in a spatially uniform continuum fault model. Geophys Res Lett 21:1983–1986
188. Shaw BE (1995) Frictional weakening and slip complexity in earthquake faults. J Geophys Res 102:18239–18251
189. Shaw BE (1997) Model quakes in the two-dimensional wave equation. J Geophys Res 100:27367–27377
190. Shcherbakov R, Turcotte DL (2004) A modified form of Bath's law. Bull Seismol Soc Am 94(5):1968–1975
191. Shnirman MG, Blanter EM (1998) Self-organized criticality in a mixed hierarchical system. Phys Rev Lett 81:5445–5448
192. Smalley RF Jr, Turcotte DL, Solla SA (1985) A renormalization group approach to the stick-slip behavior of faults. J Geophys Res 90:1894–1900
193. Sornette A, Sornette D (1989) Self-organized criticality and earthquakes. Europhys Lett 9:197–202
194. Sornette A, Sornette D (1999) Earthquake rupture as a critical point: Consequences for telluric precursors. Tectonophysics 179:327–334
195. Sornette A, Davy P, Sornette D (1990) Growth of fractal fault patterns. Phys Rev Lett 65:2266–2269
196. Sornette A, Davy P, Sornette D (1990) Fault growth in brittle-ductile experiments and the mechanics of continental collisions. J Geophys Res 98:12111–12139
197. Sornette D (1991) Self-organized criticality in plate tectonics. In: Proceedings of the NATO ASI. vol 349, "Spontaneous formation of space-time structures and criticality" Geilo, Norway 2–12 April 1991. Riste T, Sherrington D (eds) Kluwer, Dordrecht, Boston, pp 57–106
198. Sornette D (1992) Critical phase transitions made self-organized: a dynamical system feedback mechanism for self-organized criticality. J Phys I France 2:2065–2073. doi:10.1051/jp1:1992267
199. Sornette D (1998) Discrete scale invariance and complex dimensions. Phys Rep 297(5):239–270
200. Sornette D (1999) Earthquakes: from chemical alteration to mechanical rupture. Phys Rep 313(5):238–292
201. Sornette D (2000) Mechanochemistry: an hypothesis for shallow earthquakes. In: Teisseyre R, Majewski E (eds) Earthquake Thermodynamics and Phase Transformations in the Earth's Interior. Int Geophys Series, vol 76. Cambridge University Press, Cambridge, pp 329–366, e-print at http://xxx.lanl.gov/abs/cond-mat/9807400
202. Sornette D (2002) Predictability of catastrophic events: material rupture, earthquakes, turbulence, financial crashes and human birth. Proc Nat Acad Sci USA 99:2522–2529
203. Sornette D (2004) Critical Phenomena in Natural Sciences: Chaos, Fractals, Selforganization and Disorder: Concepts and Tools, 2nd edn. Springer, Berlin, p 529
204. Sornette D, Helmstetter A (2002) Occurrence of Finite-Time-Singularity in Epidemic Models of Rupture, Earthquakes and Starquakes. Phys Rev Lett 89(15):158501
205. Sornette D, Ouillon G (2005) Multifractal Scaling of Thermally-Activated Rupture Processes. Phys Rev Lett 94:038501
206. Sornette D, Pisarenko VF (2003) Fractal Plate Tectonics. Geophys Res Lett 30(3):1105. doi:10.1029/2002GL015043
207. Sornette D, Sammis CG (1995) Complex critical exponents from renormalization group theory of earthquakes: Implications for earthquake predictions. J Phys I France 5:607–619
208. Sornette D, Virieux J (1992) A theory linking large time tectonics and short time deformations of the lithosphere. Nature 357:401–403
209. Sornette D, Werner MJ (2005) Constraints on the size of the smallest triggering earthquake from the epidemic-type aftershock sequence model, Båth's law, and observed aftershock sequences. J Geophys Res 110:B08304. doi:10.1029/2004JB003535
210. Sornette D, Werner MJ (2005) Apparent clustering and apparent background earthquakes biased by undetected seismicity. J Geophys Res 110:B09303. doi:10.1029/2005JB003621
211. Sornette D, Davy P, Sornette A (1990) Structuration of the lithosphere in plate tectonics as a self-organized critical phenomenon. J Geophys Res 95:17353–17361
212. Sornette D, Vanneste C, Sornette A (1991) Dispersion of b-values in Gutenberg–Richter law as a consequence of a proposed fractal nature of continental faulting. Geophys Res Lett 18:897–900
213. Sornette D, Miltenberger P, Vanneste C (1994) Statistical physics of fault patterns self-organized by repeated earthquakes. Pure Appl Geophys 142:491–527. doi:10.1007/BF00876052

214. Sornette D, Miltenberger P, Vanneste C (1995) Statistical physics of fault patterns self-organized by repeated earthquakes: synchronization versus self-organized criticality. In: Bouwknegt P, Fendley P, Minahan J, Nemeschansky D, Pilch K, Saleur H, Warner N (eds) Recent Progresses in Statistical Mechanics and Quantum Field Theory. Proceedings of the conference 'Statistical Mechanics and Quantum Field Theory', USC, Los Angeles, May 16–21, 1994. World Scientific, Singapore, pp 313–332
215. Sornette D, Utkin S, Saichev A (2008) Solution of the Nonlinear Theory and Tests of Earthquake Recurrence Times. Phys Rev E 77:066109
216. Stanley HE (1999) Scaling, universality, and renormalization: Three pillars of modern critical phenomena. Rev Mod Phys 71(2):S358–S366. doi:10.1103/RevModPhys.71.S358
217. Sykes LR, Jaumé S (1990) Seismic activity on neighboring faults as a long-term precursor to large earthquakes in the San Francisco Bay Area. Nature 348:595–599
218. Tiampo KF, Rundle JB, Klein W (2006) Stress shadows determined from a phase dynamical measure of historic seismicity. Pure Appl Geophys 163(11–12):2407–2416
219. Turcotte DL (1986) Fractals and fragmentation. J Geophys Res 91:1921–1926
220. Turcotte DL (1997) Fractals and Chaos in Geology and Geophysics, 2nd edn. Cambridge University Press, Cambridge, p 398
221. Turcotte DL, Newman WI, Gabrielov A (2000) A statistical physics approach to earthquakes. In: Rundle JB, Turcotte DL, Klein W (eds) GeoComplexity and the Physics of Earthquake. American Geophysical Union, Washington, pp 83–96
222. Tumarkin AG, Shnirman MG (1992) Computational seismology 25:63–71
223. Vere-Jones D (1977) Statistical theories of crack propagation. Math Geol 9:455–481
224. Vere-Jones D (2005) A class of self-similar random measure. Adv Appl Probab 37(4):908–914
225. Vere-Jones D (2006) The development of statistical seismology: A personal experience. Tectonophysics 413(1–2):5–12
226. Vere-Jones D, Robinson R, Yang W (2001) Remarks on the accelerated moment release model: problems of model formulation, simulation and estimation. Geophys J Int 144:517–531. doi:10.1046/j.1365-246X.2001.01348.x
227. Voight B (1988) A method for prediction of volcanic eruptions. Nature 332:125–130
228. Voight B (1989) A relation to describe rate-dependent material failure. Science 243:200–203
229. Werner MJ (2007) On the fluctuations of seismicity and uncertainties in earthquake catalogs: Implications and methods for hypothesis testing. Ph D thesis, University of California, Los Angeles
230. Werner MJ, Sornette D (2007) Comment on "Analysis of the Spatial Distribution Between Successive Earthquakes" by Davidsen and Paczuski. [Phys Rev Lett 94:048501 (2005)]. Phys Rev Lett 99::179801
231. Werner MJ, Sornette D (2008) Magnitude Uncertainties Impact Seismic Rate Estimates, Forecasts and Predictability Experiments. J Geophys Res 113:B08302. doi:10.1029/2007JB005427
232. Wesnousky SG (1994) The Gutenberg–Richter or characteristic earthquake distribution, which is it? Bull Seismol Soc Am 84(6):1940–1959
233. Wiemer S, Katsumata K (1999) Spatial variability of seismicity parameters in aftershock zones. J Geophys Res 104:13135–13152. doi:10.1029/1999JB900032
234. Wilson K (1979) Problems in physics with many scales of length. Sci Am 241:140–157
235. Yeomans JM (1992) Statistical Mechanics of Phase Transitions. Oxford University Press Inc, New York
236. Zaliapin I, Keilis-Borok V, Ghil M (2003) A Boolean delay equation model of colliding cascades. Part I: Multiple seismic regimes. J Stat Phys 111:815–837
237. Zaliapin I, Keilis-Borok V, Ghil M (2003) A Boolean delay equation model of colliding cascades. Part II: Prediction of critical transitions. J Stat Phys 111:839–861
238. Zaliapin I, Gabrielov A, Keilis-Borok V, Wong H (2008) Clustering analysis of seismicity and aftershock identification. Phys Rev Lett 101:018501. doi:10.1103/PhysRevLett.101.018501
239. Zee A (2003) Quantum Field Theory in a Nutshell. Princeton University Press, Princeton
240. Zhuang J, Ogata Y, Vere-Jones D (2002) Stochastic declustering of space-time earthquake occurrences. J Am Stat Assoc 97:369–380
241. Zhuang J, Ogata Y, Vere-Jones D (2004) Analyzing earthquake clustering features by using stochastic reconstruction. J Geophys Res 109:B05301. doi:10.1029/2003JB002879
242. Zöller G, Hainzl S (2002) A systematic spatiotemporal test of the critical point hypothesis for large earthquakes. Geophys Rev Lett 29:53–1
243. Zöller G, Hainzl S, Kurths J (2001) Observation of growing correlation length as an indicator for critical point behavior prior to large earthquakes. J Geophys Res 106:2167–2176. doi:10.1029/2000JB900379

Seismic Wave Propagation in Media with Complex Geometries, Simulation of

HEINER IGEL[1], MARTIN KÄSER[1], MARCO STUPAZZINI[2]

[1] Department of Earth and Environmental Sciences, Ludwig-Maximilians-University, Munich, Germany

[2] Department of Structural Engineering, Politecnico di Milano, Milano, Italy

Article Outline

Glossary

Numerical methods Processes in nature are often described by partial differential equations. Finding solutions to those equations is at the heart of many studies aiming at the explanation of observed data. Simulations of realistic physical processes requires generally the use of numerical methods – a special branch of applied mathematics – that approximate the partial differential equations and allows solving them on computers. Examples are the finite-difference, finite-element, or finite-volume methods.

Spectral elements The spectral element method is an extension of the finite element method that makes use of specific basis functions describing the solutions inside each element. These basis functions (e. g., Chebyshev or Legendre polynomials) allow the interpolation of functions exactly at certain collocation points. This is often termed spectral accuracy.

Discontinuous Galerkin method The discontinuous Galerkin method is a flavor of the finite-element method that allows discontinuous behavior of the spatial or temporal fields at the element boundaries. The discontinuities – that might be small in the case of continuous physical fields such as seismic waves – then define so-called Riemann problems that can be handled using the concepts from finite-volume techniques. Therefore, the approximate solution is updated via numerical fluxes across the element boundaries.

Parallel algorithms All modern supercomputers make use of parallel architectures. This means that a large number of processors are performing (different) tasks on different data at the same time. Numerical algorithms need to be adapted to these hardware architectures by using specific programming paradigms (e. g., the message passing interface MPI). The computational efficiency of such algorithms strongly depends on the specific parallel nature of the problem to be solved, and the requirement for inter-processor communication.

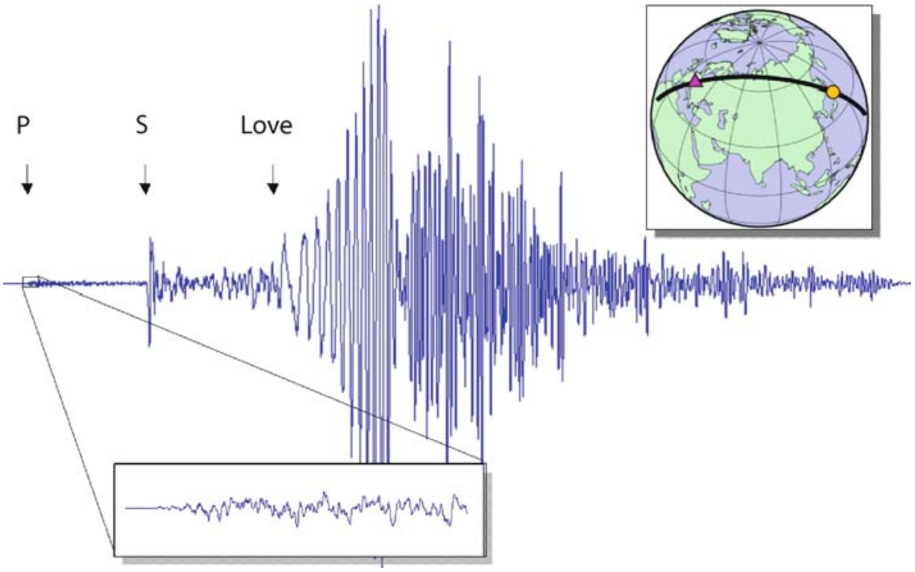
Grid generation Most numerical methods are based on the calculation of the solutions at a large set of points (grids) that are either static or depend on time (adaptive grids). These grids often need to be adapted to the specific geometrical properties of the objects to be modeled (volcano, reservoir, globe). Grids may be designed to follow domain boundaries and internal surfaces. Before specific numerical solvers are employed the grid points are usually connected to form triangles or rectangles in 2D or hexahedra or tetrahedra in 3D.

Definition of the Subject

Seismology is the science that aims at understanding the Earth's interior and its seismic sources from measurements of vibrations of the solid Earth. The resulting images of the physical properties of internal structures and the spatio-temporal behavior of earthquake rupture processes are prerequisites to understanding the dynamic evolution of our planet and the physics of earthquakes. One of the key ingredients to obtain these images is the calculation of synthetic (or theoretical) seismograms for given earthquake sources and internal structures. These synthetic seismograms can then be compared quantitatively with observations and acceptable models be searched for using the theory of inverse problems. The methodologies to calculate synthetic seismograms have evolved dramatically over the past decades in parallel with the evolution of computational resources and the ever increasing volumes of permanent seismic observations in global and regional seismic networks, volcano monitoring networks, and experimental campaigns. Today it is a tremendous challenge to extract an optimal amount of information from seismograms. The imaging process is still primarily carried out using ray theory or extensions thereof not fully taking into account the complex scattering processes that are occurring in nature.

To model seismic observations in their full complexity we need to be able to simulate wave propagation through 3D structures with constitutive relations that account for anisotropic elasticity, attenuation, porous media as well as complex internal interfaces such as layer boundaries or fault systems. This implies that numerical methods have to be employed that solve the underlying partial differential equations on computational grids. The high-frequency oscillatory nature of seismic wave fields makes this an expensive endeavor as far as computational resources are concerned. As seismic waves are propagating hundreds of wavelengths through scattering media, the required accuracy of the numerical approximations has to be of the highest possible order. Despite the fact that the physics of wave propagation is well understood, only recently computational algorithms are becoming available that allow us to accurately simulate wave propagation on many scales such as reservoirs, volcanoes, sedimentary basins, continents, and whole planets.

In addition to the imaging problem for subsurface structure and earthquake sources, the possibilities for 3D wave simulations have opened a new route to forecasting strong ground motions following large earthquakes in seismically active regions. In the absence of any hope to deterministically predict earthquakes, the calculation



Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 1
Transverse velocity seismogram of the M8.3 Tokachi-Oki earthquake near Hokkaido observed at station WET in Germany with a broadband seismometer. The total seismogram length is one hour. Arrival times of body wave phases (P, S) and the onset of transversely polarized surface (Love) waves are indicated

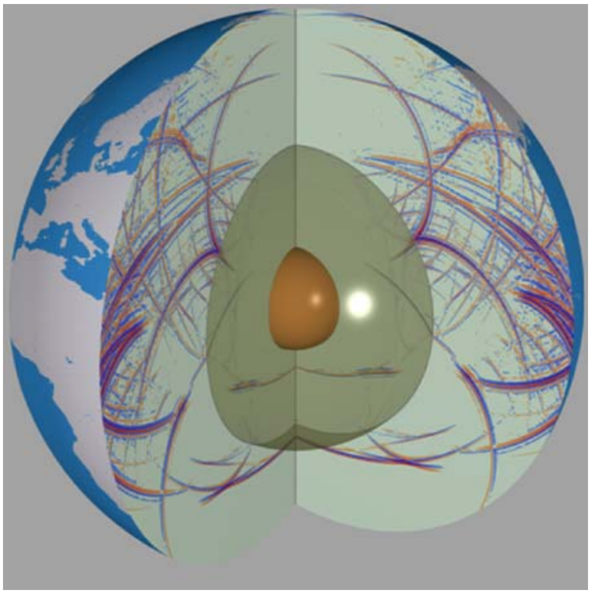
of earthquake scenarios in regions with sufficiently well known crustal structures and fault locations will play an important role in mitigating damage particularly due to potentially amplifying local velocity structures. However, to be able to employ the advanced 3D simulation technology in an efficient way, and to make use of the fast advance of supercomputing infrastructure, a paradigm shift in the concept of wave simulation software is necessary: The Earth science community has to build soft infrastructures that enable massive use of those simulation tools on the available high-performance computing infrastructure.

In this paper we want to present the state of the art of computational wave propagation and point to necessary developments in the coming years, particularly in connection with finding efficient ways to generate computational grids for models with complex topography, faults, and the combined simulation of soil and structures.

Introduction

We first illustrate the evolution of methodologies to calculate and model aspects of seismic observations for the case of global wave propagation. Seismology can look back at almost 50 years of systematic observations of earthquake induced teleseismic ground motions with the standardized global seismic and regional networks. The digital revolution in the past decades has altered the recording culture such that now seismometers are recording ground motions permanently rather than in trigger-mode, observations are becoming available in near-real time, and – because of the required sampling rates – the daily amount of observations automatically sent to the data centers is gigantic. If we take a qualitative look at a seismic observation (Fig. 1) we can illustrate what it takes to model either part or the whole information contained in such physical measurements.

In Fig. 1 a seismogram observed using a broadband seismometer (station WET in Germany) is shown. Globally observed seismograms following large earthquakes contain frequencies up to 1 Hz (P-wave motions) down to periods of around one hour (eigenmodes of the Earth) in which case modeling is carried out in the frequency domain. Seismograms of the kind shown in Fig. 1 contain many types of information. For large earthquakes the first part of the seismogram (inlet) contains valuable information on the spatio-temporal evolution of the earthquake rupture on a finite-size fault. A model of the fault slip history is a prerequisite to model the complete wave form of seismograms as the whole seismogram is affected by it unless severe low-pass filtering is applied. Information on the global seismic velocity structure is contained in the arrival times of numerous body-wave phases (here only P- and S-wave arrivals are indicated) and in the dispersive behavior of the surface waves (here the onset of the low-frequency Love waves is indicated). Further information is contained in the characteristics of the coda to body wave phases in-



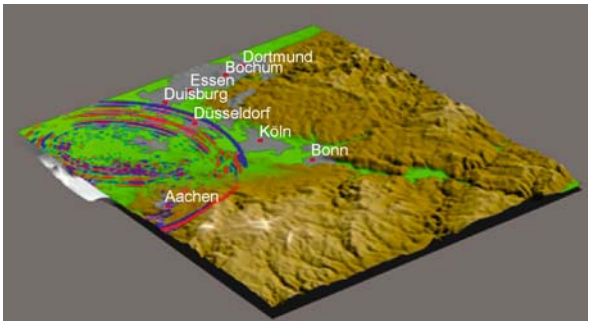
Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 2
Snapshot of wave propagation inside the Earth approx. 25 minutes after an earthquake occurs at the top part of the model. The radial component of motion is shown (blue and red denote positive and negative velocity, resp.). The simulation was carried out using an axi-symmetric approximation to the wave equation [55,58] and high-order finite-differences. Motion is allowed in the radial and horizontal directions. This corresponds to the P-SV case in 2D cartesian calculations. Therefore the wavefield contains both P- and S-waves and phase conversions

dicative of scattering in various parts of the Earth (see [62] for an account of modern observational seismology).

Adding a temporal and spatial scale to the above qualitative discussion reveals some important insight what it takes to simulate wave propagation on a planetary scale using grid-based numerical methods. Given the maximum frequency of around 1 Hz (P-waves) and 0.2 Hz (S-waves) the minimum wavelength in the Earth is expected to be O(km), requiring O(100 m) type grid spacing at least in the crustal part of the Earth leading to O(10^{12}) necessary grid points (or volume elements) for accurate numerical simulations. This would lead to memory requirements O(100 TByte) that are today possible on some of the world's largest supercomputers. The message here is that despite the rapid evolution of computational power, the complete modeling of teleseismic observations using approaches such as spectral elements (e. g., [63,64]) requiring tremendous numbers of calculations to constrain structure and sources will remain a grand challenge for some time to come. However, in many cases it is not necessary or not even desirable to simulate or model the whole seismogram, i. e. the complete observed frequency band. If we lower the cutoff frequency to 0.1 Hz (period 10 s), the required memory drops down to O(100 GByte). Such calculations can be done today on PC-clusters that can be inexpensively assembled and run on an institutional level (e. g., [8]). In addition, it means that the massive use of such forward simulations for imaging purposes and phenomenological investigations of wavefield effects is around the corner. This does not only apply to wave propagation or imaging on a planetary scale but in the same way to problems in volcanology, regional seismology, and exploration geophysics.

An illustration of global wave simulations using the finite difference method (e. g., [14,54,55,58,109,110,114]) is shown in Fig. 2 (more details on the methodologies are given in Sect. "The Evolution of Numerical Methods and Grids"). The snapshot of the radial component of motion at a time when the direct P-wave has almost crossed the Earth reveals the tremendous complexity the wave field exhibits even in the case of a spherically symmetric Earth model (PREM, [37]). The wavefield with a dominant period of ca. 15 seconds also highlights the short wavelengths that need to be propagated over very large distances. This is the special requirement for computational wave propagation that is quite different in other fields of computational Earth Sciences. While the theory of linear elastic wave propagation is well understood and most numerical methods have been applied to it in various forms, the accuracy requirements are so high that – particularly when models with complex geometrical features need to be modeled – there are still open questions as to what works best. One of the main goals of this paper is to highlight the need to focus on the grid generation process for various types of computational grid cells (e. g., rectangular, triangular in 2D, and hexahedral and tetrahedral in 3D) and the interface to appropriate highly accurate solvers for wave propagation problems.

As mentioned above computational modeling of strong ground motions following large earthquakes (see Fig. 3 for an illustration) is expected to play an increasingly important role in producing realistic estimates of shaking hazard. There are several problems that are currently unsolved: (1) to achieve frequencies that are interesting for earthquake engineers in connection with structural damage the near surface velocity structure needs to be known and frequencies beyond 5 Hz need to be calculated. In most cases this structure is not well known (on top of the uncertainties of the lower basin structures) and the required frequencies demand extremely large computational models. (2) In addition to structural uncertain-



Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 3
Snapshot (horizontal component) for a simulation of the M5.9 Roermond earthquake in the Cologne Basin in 1992 [38]. The 3D sedimentary basin (maximum depth 2 km) leads to strong amplification and prolongation of the shaking duration that correlates well with basin depth. Systematic calculations may help mitigating earthquake induced damage

ties, there are strong dependencies on the particular earthquake rupture process that influence the observed ground motions. This suggests that many 3D calculations should be carried out for any characteristic earthquake of interest, to account for such variations (e. g., Wang et al. 2006). (3) The large velocity variations (e. g., 300 m/s up to 8 km/s) require locally varying grid densities which is difficult to achieve with some of the classical numerical methods in use (e. g. finite differences). Some of the potential routes are developed below.

In summary, computational simulation of 3D wave propagation will be more and more a central tool for seismology with application in imaging problems, earthquake rupture problems, questions of shaking hazard, volcano seismology and planetary seismology. In the following we briefly review the history of the application of numerical methods to wave propagation problems and the evolution of computational grids. The increasing complexity of models in terms of geometrical features and range of physical properties imposes the use of novel methodologies that go far beyond the initial approximations based on finite differences.

The Evolution of Numerical Methods and Grids

In this section we give a brief history of the application of numerical methods to the problem of seismic wave propagation. Such a review can not be complete, certainly gives a limited perspective, and only some key references are given. One of the points we would like to highlight is the evolution of the computational grids that are being employed for wave propagation problems and the consequences on the numerical methods of choice now and in the future.

Why do we need numerical approximations to elastic wave propagation problems at all? It is remarkable what we learned about the Earth without them! In the first decades in seismology, modeling of seismic observations was restricted to the calculation of ray-theoretical travel times in spherically symmetric Earth models (e. g., [13,16]). With the advent of computing machines these approaches could be extended to 2D and 3D media leading to ray-theoretical tomography and the images of the Earth's interior that we know today (e. g., [115]). The analytical solution of wave propagation in spherical coordinates naturally leads to spherical harmonics and the possible quasi-analytical solution of wave propagation problems in spherically symmetric media using normal modes. As this methodology leads to complete waveforms the term "waveform inversion" was coined for fitting the waveforms of surface waves by correcting the phase differences for surface waves at particular frequencies (e. g., [118]). This allowed the recovery of seismic velocity models particularly of crust and upper mantle (surface wave tomography). A similar approach in Cartesian layered geometry led to complete solutions of the wave equation in cylindrical coordinates through the summation of Bessel functions, the reflectivity method [46]. This method was later extended to spherical media through the Earth-flattening transformation [85]. Recently, ray-theory was extended allowing the incorporation of finite-frequency effects (e. g., [84]). The impact on the imaging process is still being debated.

Most of these methods are still today extremely valuable in providing first estimates of 2D or 3D effects and are important for the use in standard seismic processing due to their computational efficiency. Nevertheless, with the tremendous improvements of the quality of seismic observations we strive today to extract much more information on Earth's structure and sources from recorded waveforms. As waveforms are in most places strongly affected by 3D structural variations the application of numerical methods that solve "directly" the partial differential equations descriptive of wave propagation becomes mandatory. This necessity was recognized early on and the developments of numerical wave propagation began in the sixties of the 20th century.

Numerical Methods Applied to Wave Propagation Problems

The finite-difference technique was the first numerical method to be intensively applied to the wave propagation problem (e. g., [1,6,61,77,82,83,88,89,116,117]). The partial differentials in the wave equation are replaced by finite differences leading to an extrapolation scheme in time that can either be implicit or explicit. The analysis of such simple numerical schemes led to concepts that are central to basically all numerical solutions of wave propagation problems. First, the discretization in space and time introduces a scale into the problem with the consequence that the numerical scheme becomes dispersive. This numerical dispersion – for the originally non-dispersive problem of purely elastic wave propagation – has the consequence that for long propagation distances wave pulses are no longer stable but disperse. The consequence is, that in any simulation one has to ascertain that enough grid points per wavelength are employed so that numerical dispersion is reduced sufficiently. Finding numerical schemes that minimize these effects has been at the heart of any new methodology ever since. Second, the so-called CFL criterion [24] that follows from the same theoretical analysis of the numerical scheme basically relates a "grid velocity" – the ratio between the space and time increments dx and dt, respectively – to the largest physical velocity c in the model. In order to have a stable calculation, this ratio has to be smaller than a constant ε that depends on the specific scheme and the space dimension

$$c\frac{\mathrm{d}t}{\mathrm{d}x} \leq \varepsilon\,. \tag{1}$$

This simple relationship has important consequences: When the grid spacing dx must be small, because of model areas with low seismic velocities, then the time step dt has to be made smaller accordingly leading to an overall increase in the number of time steps and thus overall computational requirements. In addition, the early implementations where based on regular rectangular grids, implying that large parts of the model where carrying out unnecessary calculations. As shown below local time-stepping and local accuracy are important ingredients in efficient modern algorithms.

The fairly inaccurate low order spatial finite-difference schemes were later extended to high-order operators [26,48,49,50,51,56,76,103]. Nevertheless, the required number of grid points per wavelength was still large, particularly for long propagation distances. This has led to the introduction of pseudo-spectral schemes, "pseudo" because only the calculations of the derivatives where done in the spectral domain, but the wave equation was still solved in the space-time domain with a time-extrapolation scheme based on finite differences [10,45,47,67]). The advantage of the calculation of derivatives in the spectral domain is at hand: The Fourier theorem tells us that by multiplying the spectrum with ik, i being the imaginary unit and k the wavenumber, we obtain an *exact* derivative (exact to numerical precision) on a regular set of grid points. This sounds attractive. However, there are always two sides to the coin. The calculation requires FFTs to be carried out extensively and the original "local" scheme becomes a "global" scheme. This implies that the derivative at a particular point in the computational grid becomes dependent on any other point in the grid. This turns out to be computationally inefficient, in particular on parallel hardware. In addition, the Fourier approximations imply periodicity which makes the implementation of boundary conditions (like the free surface, or absorbing boundary conditions) difficult.

By replacing the basis functions (Fourier series) in the classical pseudo-spectral method with Chebyshev polynomials that are defined in a limited domain (−1,1) the problem with the implementation of boundary problems found an elegant solution (e. g., [66,107,108]). However, through the irregular spacing of the Chebyshev collocation points (grid densification at the domain boundaries, see section below) new problems arose with the consequence that this approach was not much further pursued except in combination with a multi-domain approach in which the field variables exchange their values at the domain boundaries (e. g., [108]).

So far, the numerical solutions described are all based on the *strong* form of the wave equation. The finite-element method is another main scheme that found immediate applications to wave propagation problems (e. g., [79]). Finite element schemes are based on solving the *weak* form of the wave equation. This implies that the space- and time-dependent fields are replaced by weighted sums of basis (also called trial) functions defined inside elements. The main advantage of finite element schemes is that elements can have arbitrary shape (e. g., triangular, trapezoidal, hexahedral, tetrahedral, etc.). Depending on the polynomial order chosen inside the elements the spatial accuracy can be as desired. The time-extrapolation schemes are usually based on standard finite differences. There are several reasons why finite-element schemes were less widely used in the field of wave propagation. First, in the process a large system matrix needs to be assembled and must be inverted. Matrix inversion in principle requires global communication and is therefore not optimal on parallel hardware. Second, in comparison with the finite-element method, finite- difference schemes are more easily coded and implemented due to their algorithmic simplicity.

A tremendous step forward was the introduction of basis functions inside the elements that have spectral accuracy, e. g., Chebyshev or Legendre polynomials [11,15,39,40,65,86,90,98]. The so-called spectral element scheme became particularly attractive with the discovery that – by using Legendre polynomials – the matrices that required inversion became diagonal [65]. This implies that the scheme does no longer need global communication, it is a local scheme in which extrapolation to the next time step can be naturally parallelized. With the extension of this scheme to spherical grids using the cubed-sphere discretization [63,64] this scheme is today the method of choice on many scales unless highly complex models need to be initiated.

Most numerical schemes for wave propagation problems were based on regular, regular stretched, or hexahedral grids. The numerical solution to unstructured grids had much less attention, despite the fact that highly complex models with large structural heterogeneities seem to be more readily described with unstructured point clouds. Attempts were made to apply finite volume schemes to this problem [31], and other concepts (like natural neighbor coordinates [7] to find numerical operators that are applicable on unstructured grids [72,73,78]). These approaches were unfortunately not accurate enough to be relevant for 3D problems. Recently, a new flavor of numerical method found application to wave propagation on triangular or tetrahedral grids. This combination of a discontinuous Galerkin method with ideas from finite volume schemes [33,70] allows for the first time arbitrary accuracy in space and time on unstructured grids. While the numerical solution on tetrahedral grids remains computationally slower, there is a tremendous advantage in generating computational grids for complex Earth models. Details on this novel scheme are given below.

Before presenting two schemes (spectral elements and the discontinuous Galerkin method) and some applications in more detail we want to review the evolution of grids used in wave propagation problems.

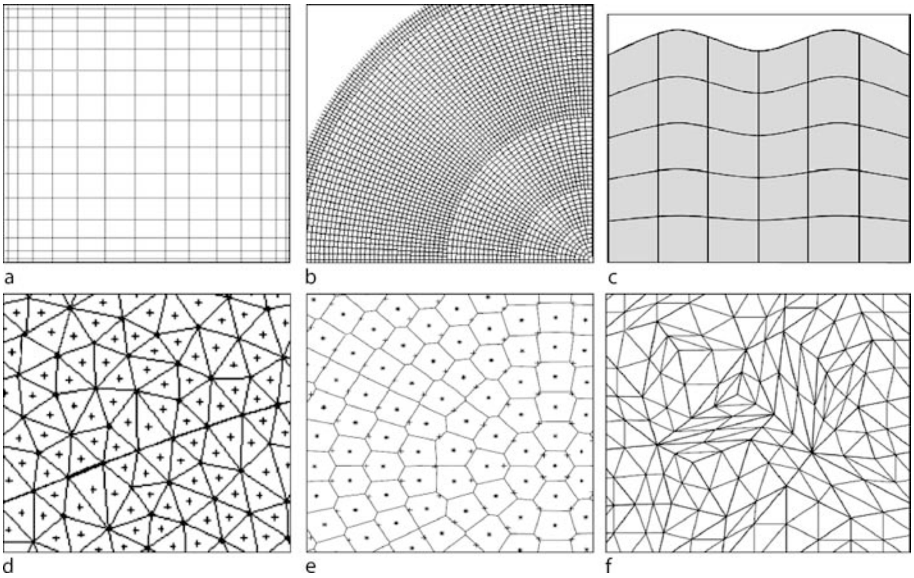
Grids for Wave Propagation Problems

The history of grid types used for problems in computational wave propagation is tightly linked to the evolution of numerical algorithms and available computational resources. The latter in the sense that – as motivated in the introduction – even today realistic simulations of wave propagation are still computationally expensive. This implies that it is not sufficient to apply stable and simple numerical schemes and just use enough grid points per wavelength and/or extremely fine grids for geometrically complex models. Optimal mathematical algorithms that minimize the computational effort are still sought for as the recent developments show that are outlined in the following sections.

In Fig. 4 a number of different computational grids in two space dimensions is illustrated. The simple-most equally-spaced regular finite-difference grid is only of practical use in situations without strong material discontinuities. With the introduction of the pseudospectral method based on Chebyshev polynomials grids as shown in Fig. 4a grids appeared that are denser near the domain boundaries and coarse in the interior. While this enabled a much more efficient implementation of boundary conditions the ratio between the size of the largest to the smallest cell depends on the overall number of grid points per dimension and can be very large. This leads to very small time steps, that can in some way be compensated by grid stretching [9] but overall the problem remains. An elegant way of allowing grids to be of more practical shape is by stretching the grids using analytical functions (Fig. 4c, this basically corresponds to a coordinate transformation, e. g., [50,107]). By doing this either smooth surface topography or smoothly varying internal interfaces can be followed by the grid allowing a more efficient simulation of geometrical features compared to a blocky representation on standard finite difference grids.

The problem of global wave propagation using spherical coordinates (here in the two-dimensional, axi-symmetric approximation) nicely illustrates the necessity to have spatially varying grid density (e. g., [42,43,53,59,89,109]). The grid shown in Fig. 4b demonstrates that in spherical coordinates a regular discretization leads to grid distances that get smaller and smaller towards the center of the Earth. This is in contrast to what is required to efficiently model the Earth's velocity structure: Velocities are small near the surface (requiring high grid density) and increase towards the center of the Earth (requiring low grid density). One way of adjusting is by re-gridding the mesh every now and then, in this case doubling the grid spacing appropriately. This is possible, yet it requires interpolation at the domain boundaries that slightly degrades the accuracy of the scheme.

The problems with grid density and complex surfaces cry for the use of so-called unstructured grids. Let us define an unstructured grid as an initial set of points (a point cloud), each point characterized by its spatial coordinates. We wish to solve our partial differential equations on this point set. It is clear that – with appropriate grid generation software – it is fairly easy to generate such grids that obey exactly any given geometrical constraints be it in connection with surfaces or velocity mod-

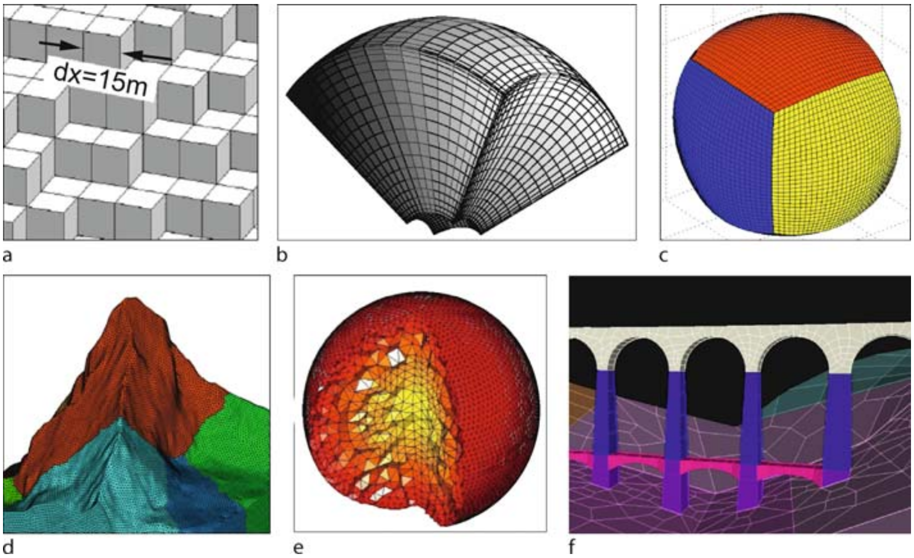


Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 4
Examples of 2D grids used for wave propagation simulations. a Chebyshev grid with grid densification near the domain boundaries. b Multidomain finite-difference grid in regular spherical coordinates. c Stretched regular finite-difference grid that allows following smoothly varying interface or surface boundaries. d Triangular staggered grid following an interface that allows finite-difference type operators. e Unstructured grid with associated Voronoi cells for calculations using the finite-volume method. f Triangular cells for finite-element type calculations. See text for details and references

els (i. e., varying grid density). It is important to note that such point clouds cannot be represented by 2D or 3D matrices as is the case for regular or regular stretched grid types. This has important consequences for the parallelization of numerical schemes. The first step after defining a point set is to use concepts from computational geometry to handle the previously unconnected points. This is done through the idea of Voronoi cells, that uniquely define triangles and their neighbors (Delauney triangulation). In Fig. 4d an example is shown for a triangular grid that follows an internal interface [72]. For finite-difference type operators on triangular grids a grid-staggering makes sense. Therefore, velocities would be defined in the center of triangles and stresses at the triangle vertices. Voronoi cells (Fig. 4e) can be used as volumetric elements for finite volume schemes [31,73]. For finite-element schemes triangular elements (Fig. 4f, e. g., [70]) with appropriate triangular shape functions are quite standard but have not found wide applications in seismology.

If the grid spacing of a regular finite-difference grid scheme in 3D would have to be halved this would result in an overall increase of computation time by a factor of 8 (a factor 2 per space dimension and another factor 2 because of the necessary halving of the time step). This simply means that the accuracy of a specific numerical scheme and the saving in memory or computation time is much more relevant in three dimensions. The evolution of grids in three dimensions is illustrated with examples in Fig. 5. A geometrical feature that needs to be modeled correctly particularly in volcanic environments is the free surface. With standard regular-spaced finite-difference schemes only a stair-step representation of the surface is possible (Fig. 5a, e. g., [87,92]). While the specific numerical implementation is stable and converges to the correct solution a tremendous number of grid points is necessary to achieve high accuracy.

Chebyshev grids and regular grids were applied to the problem of wave propagation in spherical sections (Fig. 5b, e. g., [52,57]). The advantage of solving the problem in spherical coordinates is the natural orthogonal coordinate system that facilitates the implementation of boundary conditions. However, due to the nature of spherical coordinates the physical domain should be close to the equator and geographical models have to be rotated accordingly. A highly successful concept for wave propagation in spherical media was possible through the adoption of the cubed-sphere approach in combination with spectral-elements (Fig. 5c, [63,64]). The cubed-sphere discretization is based on hexahedral grids. Towards the center of the Earth the grid spacing is altered to keep the number of elements per wavelength approximately constant.



Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 5
Examples of 3D grids. a Stair-step representation of a complex free surface with finite-difference cells. b Chebyshev grid in spherical coordinates for a spherical section. c Cubed sphere grid used for spectral-element and multi-domain Chebyshev calculations. d Tetrahedral grid of the Matterhorn. e Tetrahedral grid of the Earth's interior with the grid density tied to the velocity model. f Hexahedral grid of bridge structure and subsurface structure for spectral-element calculations. See text for details and references

Computational grids for wave propagation based on tetrahedra (Fig. 5d,e) are only recently being used for seismic wave propagation in combination with appropriate numerical algorithms such as finite volumes [34] or discontinuous Galerkin (e. g., [70]). The main advantage is that the grid generation process is greatly facilitated when using tetrahedra compared to hexahedra. Generating point clouds that follow internal velocity structures and connecting them to tetrahedra are straight forward and efficient mathematical computations. However, as described in more detail below, tetrahedral grids require more involved computations and are thus less efficient than hexahedral grids. Complex hexahedral grids – even for combined modeling of structure and soil (Fig. 5f) are possible but – at least at present – require a large amount of manual interaction during the grid generation process. It is likely that the combination of both grid types (tetrahedral in complex regions, hexahedral in less complex regions) will play an important role in future developments.

In the following we would like to present two of the most competitive schemes presently under development, (1) the spectral element method and (2) the discontinuous Galerkin approach combined with finite-volume flux schemes. The aim is to particularly illustrate the role of the grid generation process and the pros and cons of the specific methodologies.

3D Wave Propagation on Hexahedral Grids: Soil-Structure Interactions

We briefly present the spectral element method (SEM) based on Lagrange polynomials, focusing only on its main features and on its implementation for the solution of the elasto-dynamic equations. The SEM can be regarded as a generalization of the finite element method (FEM) based on the use of high order piecewise polynomial functions. The crucial aspect of the method is the capability of providing an arbitrary increase in spatial accuracy simply enhancing the algebraic degree of these functions (the spectral degree SD). On practical ground, this operation is completely transparent to the users, who limit themselves to choosing the spectral degree at runtime, leaving to the computational code the task of building up suitable quadrature points for integration and new degrees of freedom. Obviously, the increasing spectral degree implies raising the required computational effort.

On the other hand, one can also play on the grid refinement to improve the accuracy of the numerical solution, thus following the standard finite element approach. Spectral elements are therefore a so-called "$h-p$" method, where "h" refers to the grid size and "p" to the degree of polynomials. Referring to Faccioli et al. [40], Komatitsch and Vilotte [65], Chaljub et al. [15] for further details, we briefly remind in the sequel the key features of the spectral element method adopted. We start from the wave equation for the displacement $\boldsymbol{u}$:

$$\rho \frac{\partial \boldsymbol{u}^2}{\partial t^2} = \operatorname{div} \sigma_{ij}(\boldsymbol{u}) + f\,, \quad i,j = 1 \ldots \mathrm{d}(\mathrm{d} = 2,3) \quad (2)$$

where t is the time, $\rho = \rho(x)$ the material density, $f = f(x,t)$ a known body force distribution and σ_{ij} the stress tensor. Introducing Hooke's law:

$$\sigma_{ij}(\boldsymbol{u}) = \lambda \operatorname{div} \boldsymbol{u} \delta_{ij} + 2\mu \varepsilon_{ij}(\boldsymbol{u})\,, \quad (3)$$

where

$$\varepsilon_{ij}(\boldsymbol{u}) = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) \quad (4)$$

is the strain tensor, λ and μ are the Lamé coefficients, and δ_{ij} is the Kronecker symbol, i. e. $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$, otherwise.

As in the FEM approach, the dynamic equilibrium problem for the medium can be stated in the weak, or variational form, through the principle of virtual work [121] and through a suitable discretization procedure that depends on the numerical approach adopted, can be written as an ordinary differential equations system with respect to time:

$$[M]\,\ddot{\boldsymbol{U}}(t) + [K]\,\boldsymbol{U}(t) = \boldsymbol{F}(t) + \boldsymbol{T}(t) \quad (5)$$

with matrices $[M]$ and $[K]$, respectively the mass and stiffness matrices, and vectors $\boldsymbol{F}$ and $\boldsymbol{T}$ representing the contributions of external forces and traction conditions, respectively. In our SE approach, $\boldsymbol{U}$ denotes the displacement vector at the Gauss–Lobatto–Legendre (GLL) nodes, that correspond to the zeroes of the first derivatives of Legendre polynomial of degree N. The advancement of the numerical solution in time is provided by the explicit 2nd order leap-frog scheme. This scheme is conditionally stable and must satisfy the well known and already mentioned Courant–Friedrichs–Levy (CFL) condition. The key features of the SE discretization are described in the following.

Like in the FEM standard technique, the computational domain may be split into quadrilaterals in 2D or hexahedra in 3D, both the local distribution of grid points within the single element and the global mesh of all the grid points in the domain must be assigned. Many of these grid points are shared amongst several spectral elements. Each spectral element is obtained by a mapping of a master element through a suitable transformation and all computations are performed on the master element. Research is in progress regarding the introduction of triangular spectral elements [80]. The nodes within the element where displacements and spatial derivatives are computed, on which volume integrals are evaluated, are not necessarily equally spaced (similar to the Chebyshev approach in pseudospectral methods mentioned above). The interpolation of the solution within the element is done by Lagrange polynomials of suitable degree. The integration in space is done through Gauss–Lobatto–Legendre quadrature formula.

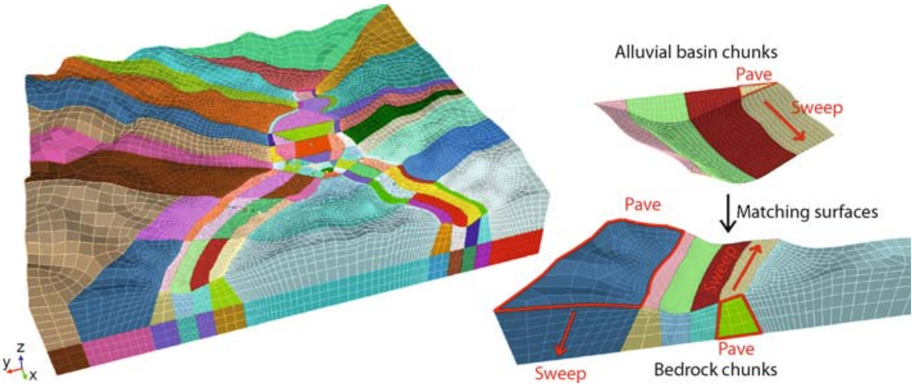
Thanks to this numerical strategy, the exponential accuracy of the method is ensured and the computational effort minimized, since the mass matrix results to be diagonal. The spectral element (SE) approach developed by Faccioli et al. [40] has been recently implemented in the computational code GeoELSE (GeoELasticity by Spectral Elements) [93,102,120] for 2D/3D wave propagation analysis. The most recent version of the code includes: (i) the capability of dealing with fully unstructured computational domains, (ii) the parallel architecture, and (iii) visco-plastic constitutive behavior [30]. The mesh can be created through an external software (e. g., CUBIT [25]) and the mesh partitioning is handled by METIS [81].

Hexahedral Grids

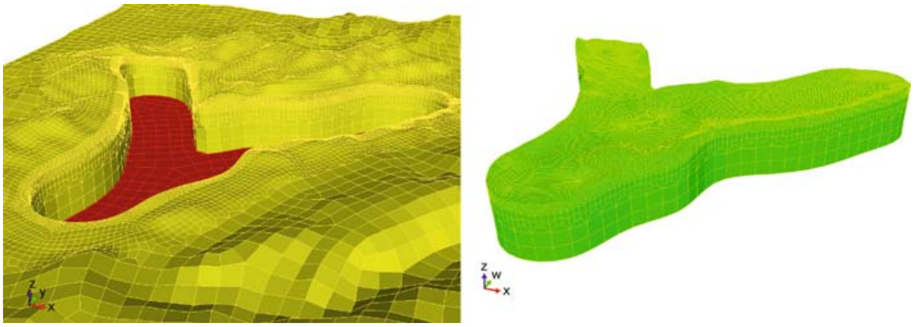
As already mentioned in the SEM here presented the computational domain is decomposed into a family of non overlapping quadrilaterals in 2D or hexahedra in 3D. The grid discretization should be suitable to accurately propagate up to certain frequencies. Obviously, owing to the strong difference of the mechanical properties between soft-soil and stiff-soil (or building construction material) and to the different geometrical details as well, the grid refinement needed in the various parts of the model varies substantially. Therefore, a highly unstructured mesh is needed to minimize the number of elements. While 3D unstructured tetrahedral meshes can be achieved quite easily with commercial or non commercial software, the creation of a 3D non structured hexahedra mesh is still recognized as a challenging problem. In the following paragraph we provide state of the art results concerning the mesh creation.

Grid Generation

Hexahedral grids have more severe restrictions in meshing efficiently. This is basically related to the intrinsic difficulty that arises from the mapping of the computational domain with this particular element. As a consequence automatic procedures have difficulty capturing specific boundaries, create poor quality elements, the assigned size is difficult to be preserved and the generation process is usually much slower compared to the tetrahedral mesh generation algo-



Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 6
3D numerical model used for the simulations of ESG06 "Grenoble Benchmark". "Honoring" technique: The computational domain is subdivided into small chunks and each one is meshed starting from the alluvial basin down to the bedrock. For simplicity only the spectral elements are shown without GLL nodes



Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 7
3D numerical model used for the simulations of ESG06 "Grenoble Benchmark". "Not Honoring" technique: The computational domain is meshed with a coarse mesh and then refined twice approximately in the area where the alluvial basin is located

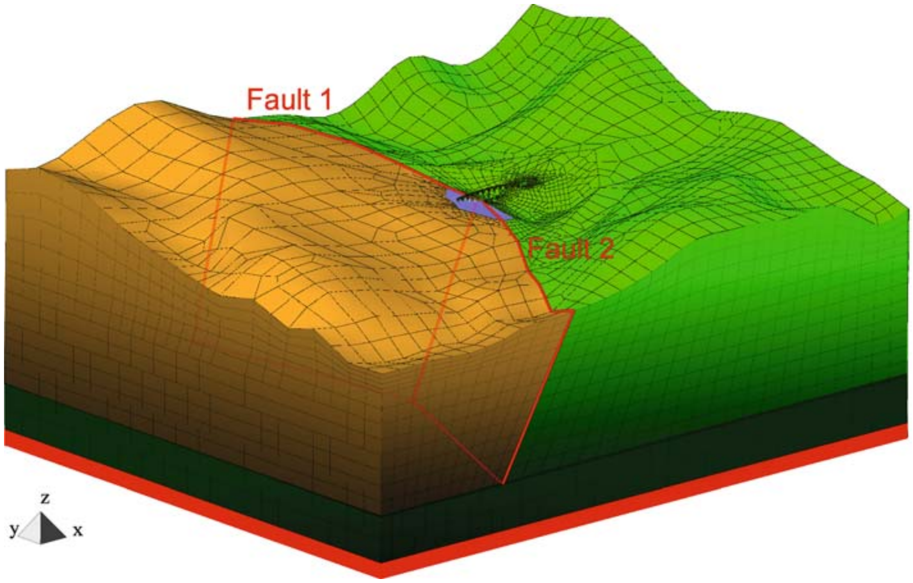
rithms. On the other hand the advantages of hexahedral meshes are usually related to the lower computational cost of the wave propagation solutions with respect to the one based on triangular meshes or hexahedral structured grids (like in the finite difference method).

Nevertheless certain problems can be addressed reasonably well with specific solutions. A quite typical case in earthquake seismology is the study of the alluvial basin response under seismic excitation. In handling this problem, a first strategy is to try to "honor" the interface between the sediment (soft soil) and the bedrock (stiff soil). The two materials are divided by a physical interface and the jump in the mechanical properties is strictly preserved. The major drawback of this approach is that usually it requires strong skills from the user to build-up the mesh and a significant amount of working time (Fig. 6). Given that the "honoring approach" is not always feasible in a reasonable time (or with a reasonable effort) a second strategy is worth to be mentioned: The so called "not honoring" procedure. In this second case the mesh is refined in proximity of the area where the soft deposit are localized but the elements do not respect the interface. On a practical ground the mechanical properties are assigned node by node and the sharp jump is smoothed through the Lagrange interpolation polynomial and substituted with smeared interfaces (Fig. 7). At the present time it is still strongly under debate if it is worth to honor or not the physical interfaces.

Finally, we highlight the fact that meshing software (e. g., CUBIT [25]) is available that seems to be extremely promising and potentially very powerful for the creation of geophysical and seismic engineering unstructured hexahedral meshes. Further very interesting mesh generation procedures based on hexahedral are under investigation [99].

Scale Problem with Structure and Soil

In engineering practice one of the most common approaches to design buildings under seismic load is the im-

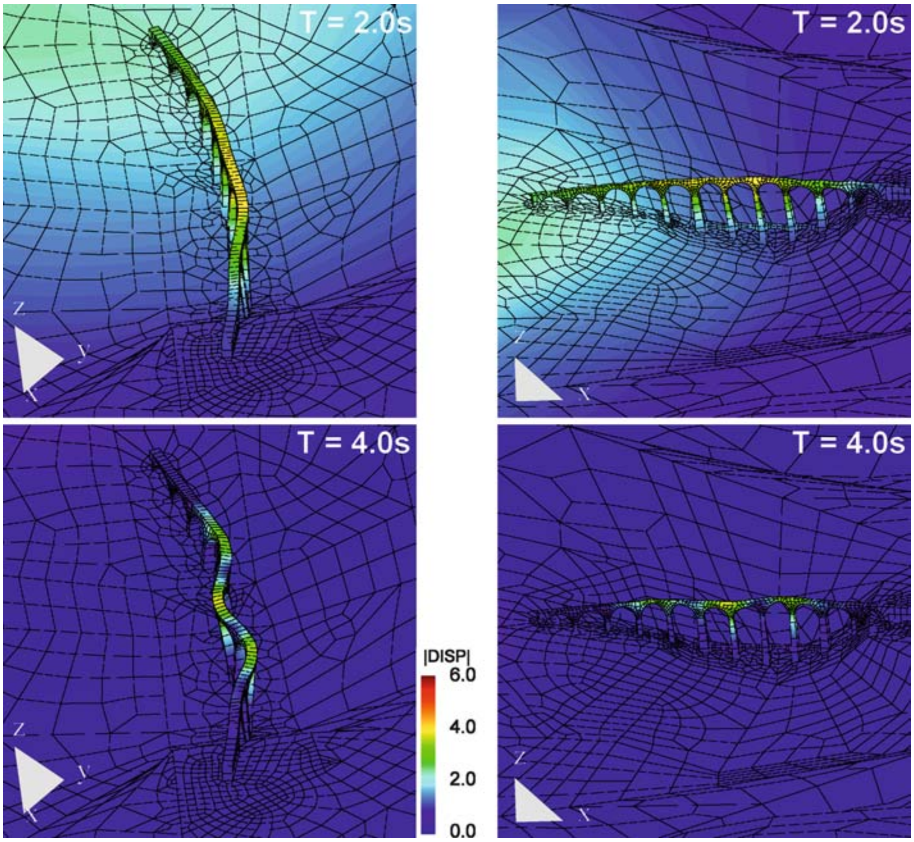


Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 8
3D model of Acquasanta bridge and the surrounding geological configuration. The investigated area is 2 km in length, 1.75 km in width and 0.86 km in depth. The model was designed to propagate waves up to 5 Hz with a SD = 3 (Order 4) and has 38,569 hexahedral elements and 1,075,276 grid points. The contact between calcareous schists (*brown color*) and serpentine rocks (*green color*) is modeled with two sub-vertical faults (*red-line*). *Cyan color* represents the alluvial and weathered deposits

position of an acceleration time history to the structure, basically acting like an external load. An excellent example of this technique can be found in recent publications (e. g., [68,69]) and in the study of the so-called "urban-seismology", recently presented by Fernandez-Ares et al., [44]. In this case the goal is to understand how the presence of an entire city can modify the incident wave-field. Due to the size of the simulation and the number of buildings, the latter are modeled as single degrees of freedom oscillators. The interaction between soil and structure is preserved but the buildings are simplified. For important structure (e. g.: Historical buildings, world heritage buildings, hospitals, schools, theaters, railway and highways) it is worth to provide an ad-hoc analysis capable to take into account the full complexity of the phenomena.

Here we present an example of a fully coupled modeling (Fig. 8): A railway bridge and its geotechnical-topographical surroundings. The Acquasanta bridge on the Genoa-Ovada railway, North Italy, is located in the Genoa district and represents a typical structure the ancestor of which can be traced back to the Roman "Pont du Gard". This structural type did not change significantly along the centuries, thanks to the excellent design achieved no less than 1900 years ago. The Acquasanta bridge structure is remarkable both for the site features and the local geological and geomorphological conditions. The foundations of several of the piers rest on weak rock; moreover, some instability problems have been detected in the past on the valley slope towards Ovada.

Several simulations have been performed with GeoELSE, in order to evaluate the influence of seismic site effects on the dynamic response of the Acquasanta bridge. A fully coupled 3D soil-structure model was designed: The grid is characterized by a "subvertical fault" between calcareous schists and serpentine rocks. This is in accordance with available data, even if further investigations in future should identify more in detail the tectonic structure of the area. The geometry of weathered materials overlaying the calcareous schists on the Ovada side has been assumed according to available information. The dimension of hexahedral elements ranges some tens of centimeters to about 1000 m. With such a model, the problem can be handled in its 3D complexity and we can examine the following aspects that are usually analyzed under restrictive and simplified assumptions: (i) soil-structure interaction, (ii) topographic amplification, (iii) soft soil amplification (caused by the superficial alluvium deposit shown in cyan), (iv) subvertical fault (red line) between the schists, on the Ovada side, and serpentine rock, on the Genoa side. For excitation a shear plane wave (x-direction) was used (Ricker wavelet, fmax = 3 Hz, $t0 = 1.0\,s$. and amplitude = 1 mm) propagating vertically from the bottom (red elements in Fig. 8).



Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 9
Snapshots of the modulus of the displacement vector and the magnified deformed shape of the bridge (in mm)

In Fig. 9 we present some snapshots of the modulus of the displacement vector and the magnified deformed shape of the bridge. It is worth to note that at $T = 2$ s the motion of the bridge is almost in-plane (direction x), while at $T = 4$ s is clearly evident how the coupling between the in-plane and out-plane (y-direction) motion starts to be important.

The study of the soil-structure interaction problem could be easily enhanced (i) improving the input excitation of the model here presented and (ii) taking into account complex constitutive behavior both from the soil and the structure side. The former is already available in GeoELSE thanks to the recent implementation [41,93] of the domain reduction method (DRM), a methodology that divides the original problem into two simpler ones [4,119], to overcome the problem of multiple physical scales that is created by a seismic source located some kilometers away from the structure with typical element size of the order of meters and located over a relatively small area (less than 1 km^2) on soft deposit. The latter still needs to be improved because of the lack of a complete tool capable to handle in 3D non linear soil behavior, non-linear structural behavior and the presence of the water, that play a crucial role in the failure of buildings. Partial response to this problem can be found in the recent work of Bonilla et al. [5] and in the visco-plastic rheology recently introduced in GeoELSE [30].

3D Wave Propagation on Tetrahedral Grids: Application to Volcanology

As indicated above, the simulation of a complete, highly accurate wave field in realistic media with complex geometry is still a great challenge. Therefore, in the last years a new, highly flexible and powerful simulation method has been developed that combines the Discontinuous Galerkin (DG) Method with a time integration method using Arbitrary high order DERivatives (ADER) of the approximation polynomials. The unique property of this numerical scheme is, that it achieves arbitrarily high approximation order for the solution of the governing seismic wave equation in space and time on structured and unstructured meshes in two and three space dimensions.

Originally, this new ADER-DG approach [32,35] was introduced for general linear hyperbolic equation systems with constant coefficients or for linear systems with variable coefficients in conservative form. Then, the extension to non-conservative systems with variable coefficients and source terms and its particular application to the simulation of seismic waves on unstructured triangular meshes in two space dimensions was presented [70]. And finally, the further extension of this approach to three-dimensional tetrahedral meshes has been achieved [33]. Furthermore, the accurate treatment of viscoelastic attenuation, anisotropy and poroelasticity has been included to handle more complex rheologies [28,29,71]. The governing system of the three-dimensional seismic wave equations is hereby formulated in velocity-stress and leads to the hyperbolic system of the form

$$\frac{\partial \boldsymbol{Q}_p}{\partial t} + A_{pq}\frac{\partial \boldsymbol{Q}_q}{\partial \xi} + B_{pq}\frac{\partial \boldsymbol{Q}_q}{\partial \eta} + C_{pq}\frac{\partial \boldsymbol{Q}_q}{\partial \zeta} = S_p \,, \qquad (6)$$

where the vector $\boldsymbol{Q}$ of unknowns contains the six stress and the three velocity components and S is the source term. The Jacobian matrices A, B and C include the material values as explained in detail in [33,70].

The ADER-DG Method: Basic Concepts

The ADER-DG method is based on the combination of the ADER time integration approach [113], originally developed in the finite volume (FV) framework [96,97,111] and the Discontinuous Galerkin finite element method [18,19,20,21,22,23,91]. As described in detail in [33] in the ADER-DG approach the solution is approximated inside each tetrahedron by a linear combination of space-dependent polynomial basis functions and time-dependent degrees of freedom as expressed through

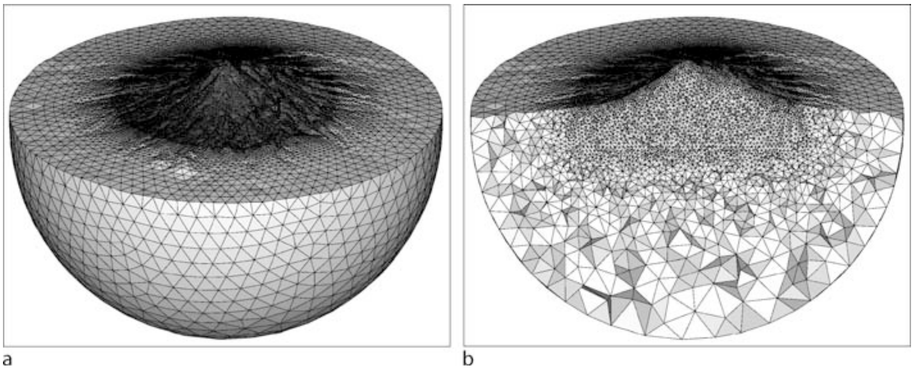
$$(\boldsymbol{Q}_h)_p(\xi,\eta,\zeta,t) = \hat{\boldsymbol{Q}}_{pl}(t)\boldsymbol{\Phi}_l(\xi,\eta,\zeta) \,, \qquad (7)$$

where the basis functions Φ_l form an orthogonal basis and are defined on the canonical reference tetrahedron. The unknown solution inside each element is then approximated by a polynomial, whose coefficients – the degrees of freedom $\hat{\boldsymbol{Q}}_{pl}$ – are advanced in time. Hereby, the solution can be discontinuous across the element interfaces, which allows the incorporation of the well-established ideas of numerical flux functions from the finite volume framework [75,112]. To define a suitable flux over the element surfaces, so-called Generalized Riemann Problems (GRP) are solved at the element interfaces. The GRP solution provides simultaneously a numerical flux function as well as a time-integration method. The main idea is a Taylor expansion in time in which all time derivatives are replaced by space derivatives using the so-called Cauchy–Kovalewski procedure which makes recursive use of the governing differential Eq. (6). The numerical solution of Eq. (6) can thus be advanced by one time step without intermediate stages as typical e. g. for classical Runge–Kutta time stepping schemes. Due to the ADER time integration technique the same approximation order in space and time is achieved automatically. Furthermore, the projection of the elements in physical space onto a canonical reference element allows for an efficient implementation, as many computations of three-dimensional integrals can be carried out analytically beforehand. Based on a numerical convergence analysis this new scheme provides arbitrary high order accuracy on unstructured meshes. Moreover, due to the choice of the basis functions in Eq. (7) for the piecewise polynomial approximation [23], the ADER-DG method shows even spectral convergence.

Grid Generation: Unstructured Triangulations and Tetrahedralization

Both tetrahedral and hexahedral elements are effectively used to discretize three-dimensional computational domains and model wave propagation with finite element type methods. Tetrahedra can be the right choice because of the robustness when meshing any general shape. Hexahedra can be the element of choice due to their ability to provide more efficiency and accuracy in the computational process. Furthermore, techniques for automatic mesh generation, gradual mesh refinement and coarsening are generally much more robust for tetrahedral meshes in comparison to hexahedral meshes. Straightforward tetrahedral refinement schemes, based on longest-edge division, as well as the extension to adaptive refinement or coarsening procedures of a refined mesh exist [3,12]. In addition, parallel strategies for refinement and coarsening of tetrahedral meshes have been developed [27].

Less attention has been given to the modification of hexahedral meshes. Methods using iterative octrees have been proposed [74,95], but these methods often result in nonconformal elements that cannot be accommodated by some solvers. Lately also conformal refinement and coarsening strategies for hexahedral meshes have been proposed [2]. Other techniques insert non-hexahedral elements that result in hybrid meshes that need special solvers that can handle different mesh topologies. Commonly, the geometrical problems in geosciences arise through rough surface topography, as shown for the Mer-



Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 10
Tetrahedral mesh for the model of the volcano Merapi. The zone of interest, such as the free surface topography and the volcano's interior are discretized by a fine mesh, whereas the spatial mesh is gradually coarsened towards the model boundaries

api volcano in Fig. 10, and internal material boundaries of complex shape that lead to wedges and overturned or discontinuous surfaces due to folding and faulting. However, once the geometry of the problem is defined by the help of modern computer aided design (CAD) software, the meshing process using tetrahedral elements is automatic and stable. After the mesh generation process, the mesh vertices, the connectivity matrix and particular information about boundary surfaces are typically imported to a solver.

The computational possibilities and algorithmic flexibility of a particular solver using the ADER-DG approach for tetrahedral meshes are presented in the following.

Local Accuracy: *p*-Adaptation

In many large scale applications the computational domain is much larger than the particular zone of interest. Often such an enlarged domain is chosen to avoid effects from the boundaries that can pollute the seismic wave field with possible, spurious reflections. Therefore, a greater number of elements has to be used to discretize the domain describing the entire model. However, in most cases the high order accuracy is only required in a restricted area of the computational domain and it is desirable to choose the accuracy that locally varies in space. This means, that it must be possible to vary the degree *p* of the approximation polynomials locally from one element to the other [36]. As the ADER-DG method uses a hierarchical order of the basis functions to construct the approximation polynomials, the corresponding polynomial coefficients, i. e. the degrees of freedom, for a lower order polynomial are always a subset of those of a higher-order one. Therefore, the computation of fluxes between elements of different approximation orders can be carried out by using only the necessary part of the flux matrices.

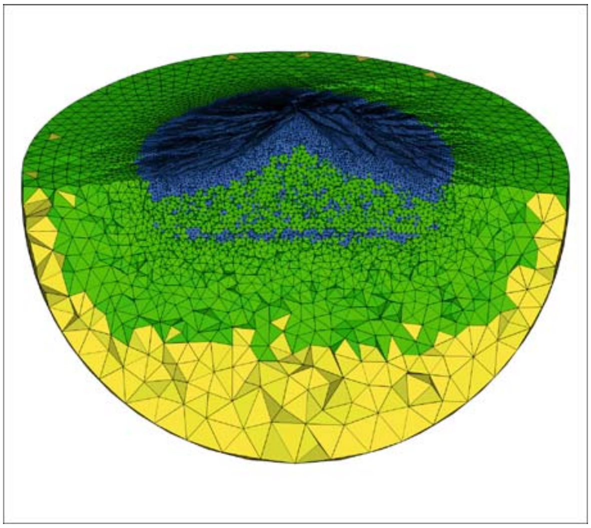
Furthermore, the direct coupling of the time and space accuracy via the ADER approach automatically leads to a local adaptation also in time accuracy, which often is referred to as p_t-adaptivity. In general, the distribution of the degree *p* might be connected to the mesh size *h*, i. e. the radius of the inscribed sphere of a tetrahedral element. In particular, the local degree *p* can be coupled to the mesh size *h* via the relations

$$p = p_{\min} + (p_{\max} - p_{\min}) \left(\frac{h - h_{\min}}{h_{\max} - h_{\min}} \right)^r , \qquad (8)$$

$$p = p_{\max} - (p_{\max} - p_{\min}) \left(\frac{h - h_{\min}}{h_{\max} - h_{\min}} \right)^r , \qquad (9)$$

where the choice of the power *r* determines the shape of the *p*-distribution. Note, that depending on the choice of the first term and the sign the degree *p* can increase as in Eq. (8) or decrease as in Eq. (9) with increasing *h*, starting from a minimum degree $p_{\min}$ up to a maximum degree $p_{\max}$. This provides additional flexibility for the distribution of *p* inside the computational domain. An example of a *p*-distribution for the volcano Merapi is given in Fig. 11.

Here the idea is to resolve the slowly propagating surface waves with high accuracy, whereas the waves propagating towards the absorbing model boundaries pass through a zone of low spatial resolution. This approach leads to numerical damping due to an amplitude decay that reduces possible boundary reflections. Furthermore, the computational cost is reduced significantly due to the strongly reduced number of total degrees of freedom in the model.



Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 11

The local degree p of the approximation polynomial depends on the insphere radius of each tetrahedral element and is given in color code. Close to the surface topography an approximation polynomial of degree $p = 5$ (*blue*) is used, whereas in depth the degree is reduced to $p = 4$ (*green*) and $p = 3$ (*yellow*)

Local Time Stepping: Δt-Adaptation

Geometrically complex computational domains or spatial resolution requirements often lead to meshes with small or even degenerate elements. Therefore, the time step for explicit numerical schemes is restricted by the ratio of the size h of the smallest element and the corresponding maximum wave speed in this element. For global time stepping schemes all elements are updated with this extremely restrictive time step length leading to a large amount of iterations. With the ADER-DG approach, time accurate local time stepping can be used, such that each element is updated by its own, optimal time step [36]. Local time-stepping was used in combination with the finite-difference method [42,106].

An element can be updated to the next time level if its actual time level and its local time step Δt fulfill the following condition with respect to all neighboring tetrahedra n:

$$t + \Delta t \leq \min(t_n + \Delta t_n) \, . \qquad (10)$$

Figure 12 is visualizing the evolution of four elements (I, II, III and IV) in time using the suggested local time stepping scheme. A loop cycles over all elements and checks for each element, if condition (10) is fulfilled. At the initial state all elements are at the same time level, however, element II and IV fulfill condition (10) and therefore can be updated. In the next cycle, these elements are already advanced in time (grey shaded) in cycle 1. Now elements I and IV fulfill condition (10) and can be updated to their next local time level in cycle 2. This procedure continues and it is obvious, that the small element IV has to be updated more frequently than the others. A synchronization to a common global time level is only necessary, when data output at a particular time level is required as shown in Fig. 12.

Information exchange between elements across interfaces appears when numerical fluxes are calculated. These fluxes depend on the length of the local time interval over which a flux is integrated and the corresponding element is evolved in time. Therefore, when the update criterion (10) is fulfilled for an element, the flux between the element itself and its neighbor n has to be computed over the local time interval:

$$\tau n = [\max(t, tn) \, , \quad \min(t + \Delta t, tn + \Delta tn)] \, . \qquad (11)$$

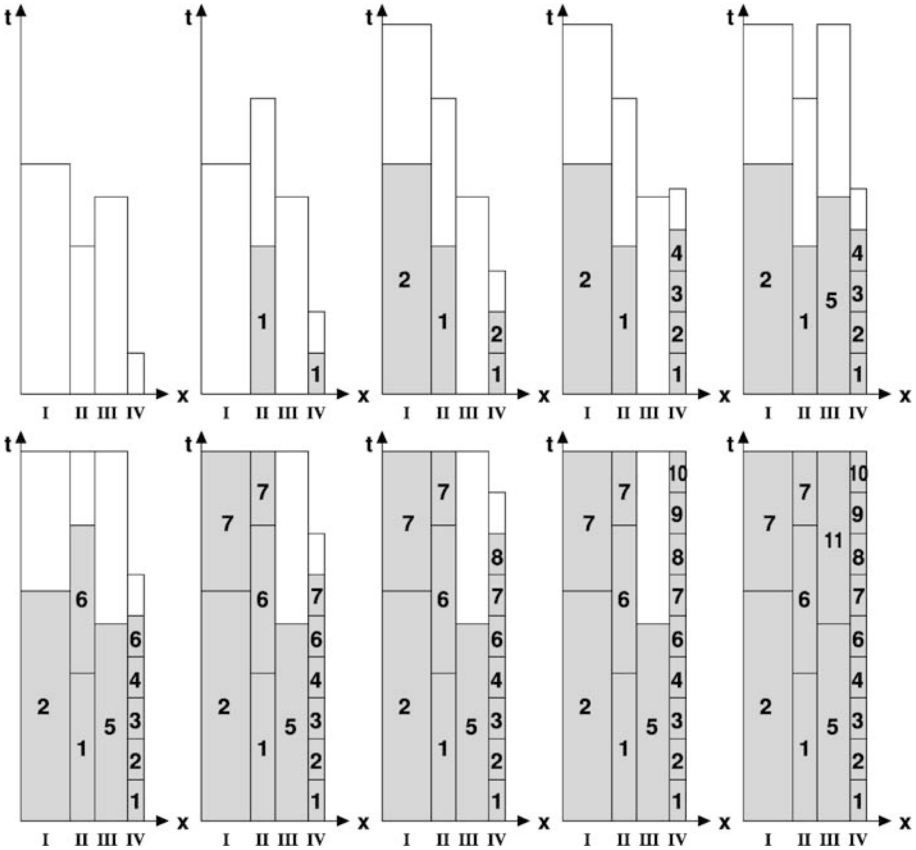
As example, the element III fulfills the update criterion (10) in cycle 5 (see Fig. 12). Therefore, when computing the fluxes only the remaining part of the flux given by the intervals in Eq. (11) has to be calculated. The other flux contribution was already computed by the neighbors II and IV during their previous local updates. These flux contributions have been accumulated and were stored into a memory variable and therefore just have to be added.

Note that e. g. element IV reaches the output time after 10 cycles and 9 local updates, which for a global time stepping scheme would require $9 \times 4 = 36$ updates for the all four elements. With the proposed local time stepping scheme only 16 updates are necessary to reach the same output time with all elements as indicated by the final number of grey shaded space time elements in Fig. 12.

Comparing these numbers leads to a speedup factor of 2.25. For strongly heterogeneous models and local time step lengths this factor can become even more pronounced. However, due to the asynchronous update of elements that might be spatially very close to each other the mesh partitioning for parallel computations becomes an important and difficult issue. Achieving a satisfying load balancing is a non-trivial task and still poses some unresolved problems as explained in the following.

Mesh Partitioning and Load Balancing

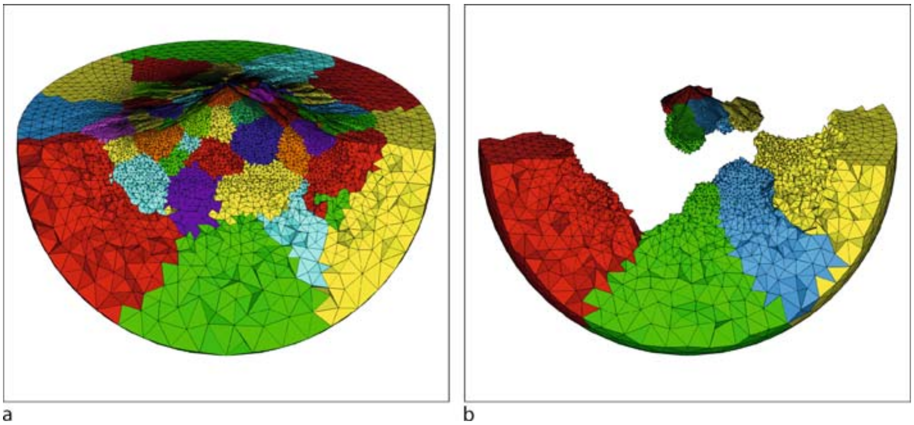
For large scale applications it is essential to design a parallel code that can be run on massively parallel supercomputing facilities. Therefore, the load balancing is an important issue to use the available computational resources efficiently. For global time stepping schemes with-



Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 12
Visualization of the local time stepping scheme. The actual local time level *t* is at the top of the *gray shaded* area with numbers indicating the cycle, in which the update was done. *Dotted lines* indicate the local time step length Δ*t* with which an element is updated

out p-adaptation standard mesh partitioning as done e. g. by METIS [60] is sufficient to get satisfying load balancing. The unstructured tetrahedral mesh is partitioned into subdomains that contain an equal or at least very similar number of elements as shown in Fig. 13. Therefore, each processor has to carry out a similar amount of calculations. However, if p-adaptation is applied, the partitioning is more sophisticated as one subdomain might have many elements of high order polynomials whereas another might have the same number of elements but with lower order polynomials. Therefore, the parallel efficiency is restricted by the processor with the highest work load. However, this problem can usually be solved by weighted partitioning algorithms, e. g. METIS.

In the case of local time stepping, mesh partitioning is becoming a much more difficult task. One solution is to divide the computational domain into a number of zones, that usually contain a geometrical body or a geological zone that typically is meshed individually with a particular mesh spacing h and contains a dominant polynomial order. Then each of these zones is partitioned separately into subdomains of approximately equal numbers of elements. Then each processor receives a subdomain of each zone, which requires a similar amount of computational work as shown in Fig. 13. In particular, the equal distribution of tetraheda with different sizes is essential in combination with the local time stepping technique. Only if each processor receives subdomains with a similar amount of small and large elements, the work load is balanced. The large elements have to be updated less frequently than the smaller elements and therefore are computationally cheaper. Note, that the separately partitioned and afterwards merged zones lead to non-connected subdomains for each processor (see Fig. 13). This increases the number of element surfaces between subdomains of different processors and therefore increases the communication required. However, communication is typically low as the degrees of freedom have to be exchanged only once per



Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 13
Standard partitioning of the computational domain (*left*) and an example of 4 subdomains grouped together for more efficient local time stepping

time step and only for tetrahedra that have an interface at the boundary between subdomains. Therefore, the improvements due to the new load balancing approach are dominant and outweigh the increase in communication.

However, care has to be taken as the distribution of the polynomial degree *p* or the seismic velocity structure might influence the efficiency of this grouped partitioning technique. A profound and thorough mesh partitioning method is still a pending task as the combination of local time stepping and *p*-adaptivity requires a new weighting strategy of the computational cost for each tetrahedral element considering also the asynchronous element update. The automatic partitioning of unstructured meshes with such heterogeneous properties together with the constraint of keeping the subdomains as compact as possible to avoid further increase of communication is still subject to future work.

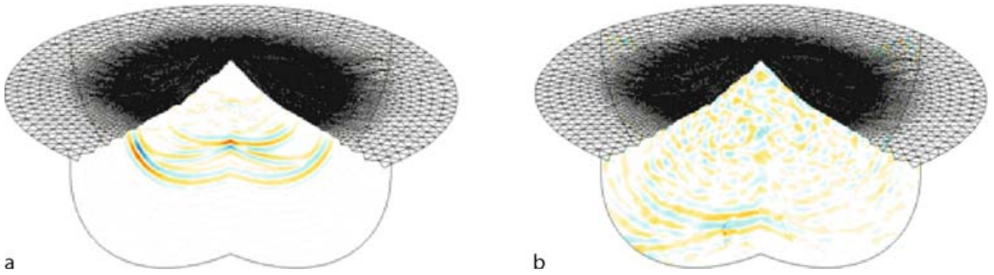
In Fig. 13 an example of a grouped partition of the tetrahedral mesh is shown for 4 processors. Two non-connected subdomains indicated by the same color are assigned to each processor including small – and therefore computationally expensive – tetrahedra that are updated frequently due to their small time step, and much larger elements that typically are cheap due to their large time step. This way, the work load often is balanced sufficiently well over the different processors.

Relevance of High Performance Computing: Application to Merapi Volcano

In recent years the development of the ADER-DG algorithm including the high order numerical approximation in space and time, the mesh generation, mesh adaptation, parameterization, and data visualization created the basis of an efficient and highly accurate seismic simulation tool. Realistic large scale applications and their specific requirements will further guide these developments. On the other hand, the study and incorporation of geophysical processes that govern seismic wave propagation insures, that the simulation technology matches the needs and addresses latest challenges in modern computational seismology. Hereby, the accurate modeling of different source mechanisms as well as the correct treatment of realistic material properties like anelasticity, viscoplasticity, porosity and highly heterogeneous, scattering media will play an important role.

However, only the combination of this state-of-the-art simulation technology with the most powerful supercomputing facilities actually available can provide excellent conditions to achieve scientific progress for realistic, large scale applications. This combination of modern technologies will substantially contribute to resolve current problems, not only in numerical seismology, but will also influence other disciplines. The phenomenon of acoustic, elastic or seismic wave propagation is encountered in many different fields. Beginning with the classical geophysical sciences seismology, oceanography, and volcanology such waves also appear in environmental geophysics, atmospheric physics, fluid dynamics, exploration geophysics, aerospace engineering or even medicine.

With the rapid development of modern computer technology and the development of new highly accurate simulation algorithms computer modeling just started to herald a new era in many applied sciences. The 3D wave propagation simulations in realistic media require a substantial amount of computation time even on large par-



Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 14
Snapshots of the seismic wave field after an explosive event close to the summit of Merapi volcano. The free surface topography introduces strong scattering of the waves making it extremely difficult to invert for the seismic source mechanism or the exact source location

allel computers. Extremely powerful national supercomputers already allow us to run simulations with unrivaled accuracy and resolution. However, using the extremely high accuracy and flexibility of new simulation methods on such massively parallel machines the professional support of experts in supercomputing is absolutely essential. Only professional porting, specific CPU-time and storage optimizations of current software with respect to continuously changing compilers, operating systems, hardware architectures or simply personnel, will ensure the lifetime of new simulation technologies accompanied by ongoing improvements and further developments. Additionally, the expertise and support in the visualization of scientific results using technologies of Virtual Reality for full 3D models not only enhances the value of simulation results but will support data interpretation and awake great interest in the new technology within a wide research community.

As an example, volcano monitoring plays an increasingly important role in hazard estimation in many densely populated areas in the world. Highly accurate computer modeling today is a key issue to understand the processes and driving forces that can lead to dome building, eruptions or pyroclastic flows. However, data of seismic observations at volcanoes are often very difficult to interpret. Inverting for the source mechanism, i. e. seismic moment tensor inversion, or just locating an exact source position is often impossible due to the strongly scattered wave field caused by an extremely heterogeneous material distribution inside the volcano. Furthermore, the rough topography alone can affect the wave field by its strongly scattering properties as shown in Fig. 14.

Therefore, it is fundamental to understand the effects of topography and scattering media and there influence on the seismic wave field. A systematic study of a large number of scenarios computed by highly accurate simulation methods to provide reliable synthetic data sets is necessary to test the capabilities of currently used inversion tools. Slight changes in parameters like the source position, the source mechanism or the elastic and geometric properties of the medium can then reveal the limits of such tools and provide more precise bounds of their applicability in volcano seismology.

Finally, the implementation of the ADER-DG method is still much more expensive than other state-of-the-art implementations of existing methods. However, a fair comparison between accuracy and computational cost is still a pending task. The main reason for the CPU-time difference is the much larger number of tetrahedral elements than hexahedra that have to be used to cover the same volume. Furthermore, due to the choice of the basis functions, the flux computations are expensive, as the matrix-matrix multiplications involved are not sparse.

However, the ADER-DG method is currently implemented on hexahedral meshes to make fair comparisons possible. Preliminary tests show, that the change of mesh topology from tetrahera to hexahedra significantly reduces the computational cost. However, final results are subject to future investigations.

Discussion and Future Directions

As indicated in the introduction and highlighted in the previous sections, computational tools for wave propagation problems are getting increasingly sophisticated to meet the needs of current scientific problems. We are far away from simple finite-difference time schemes that are solving problems on regular grids on serial computers in which case the particular programming approach did not affect dramatically the overall performance. Today, competitive algorithms are results of years of partly highly professional coding. Implementations on high-performance computing hardware requires in-depth knowledge of parallel algorithms, profiling, and many technical aspects of modern computations. To make complex scientific software available to other researchers requires implementation and testing on many different (parallel) platforms. This may involve parallelization using different programming paradigms (e. g., the combination of OpenMP and MPI on nodes of shared memory machines), and interoperability on heterogeneous computational GRIDs.

This has dramatic consequences particularly for young researchers in the Earth Sciences who want to use advanced computational tools to model observations. While in the early days a finite-difference type algorithm could be understood, coded, implemented and tested in a few weeks, this is no longer possible. In addition, standard curricula do not offer training in computational methods allowing them to efficiently write and test codes. This suggests that at least for some, well-defined computational problems verified and professionally engineered scientific software solutions should be provided to the community and also professionally extended and maintained in close collaboration with scientists. In seismology we are in a quite fortunate situation. In contrast to many other fields of physical sciences, our constitutive relations (e. g., stress-strain) are fairly well understood, and – as indicated in this paper – numerical solutions for 3D problems and their implementation on parallel hardware are well advanced. Another argument for stable tested "community"-codes for wave propagation is the fact that advancement in many scientific problems (e. g., imaging the Earth's interior, quantifying earthquake-induced shaking hazard) relies on zillions of forward modeling runs with only slight variations of the internal velocity models.

As far as technical developments are concerned, the efficient initialization of complex 3D models on computational grids is still a great challenge. Realistic models may be composed of complex topography, families of overlapping fault surfaces, discontinuous interfaces, and varying rheologies (e. g., elastic, anisotropic, viscoelastic, viscoplastic, porous). This may require the combination of tetrahedral and hexahedral grid in models with strongly varying degree of complexity. Ideally, standards for Earth models (and synthetic data) formats should be established by the communities that allow easy exchange and multiple use of models with different simulation tools (e. g., wave propagation, deformation, earthquake rupture). In addition, the rapid developments towards PetaFlop computing opens new questions about the scalability and efficient parallelization of current and future algorithms.

As the forward problem of wave propagation is at the core of the seismic imaging problem for both source and Earth's structure, in the near future we will see the incorporation of 3D simulation technology into the imaging process. Provided that the background seismic velocity models are fairly well known (e. g., reservoirs, global Earth, sedimentary basins), adjoint methods provide a powerful analytical tool to (1) relate model deficiencies to misfit in observations and (2) quantify the sensitivities to specific aspects of the observations (e. g., [100,104,105]). As the core of the adjoint calculations is the seismic forward problem, the challenge is the actual application to real data and the appropriate parametrizations of the model and the data that optimize the data fitting process.

In summary, while we look back at (and forward to) exciting developments in computational seismology, a paradigm shift in the conception of one of the central tools of seismology – the calculation of 3D synthetic seismograms – is necessary. To extract a maximum amount of information from our high-quality observations scientists should have access to high-quality simulation tools. It is time to accept that "*software is infrastructure*" and provide the means to professionally develop and maintain community codes and model libraries at least for basic Earth science problems and specific focus regions. Developments are one the way along those lines in the SPICE project (Seismic Wave Propagation and Imaging in Complex Media, a European Network [101]), the Southern California Earthquake Center (SCEC [94]) and the CIG Project (Computational infrastructure in geodynamics [17]).

Acknowledgments

We would like to acknowledge partial support towards this research from: The European Human Resources and Mobility Program (SPICE-Network), the German Research Foundation (Emmy Noether-Programme), the Bavarian Government (KOHNWIHR, graduate college THESIS, BaCaTec), and MunichRe. We would also like to thank J. Tromp supporting MS's visit to CalTech. We also thank two anonymous reviewers for constructive comments on the manuscript.

Bibliography

Primary Literature

1. Alterman Z, Karal FC (1968) Propagation of elastic waves in layered media by finite-difference methods. Bull Seism Soc Am 58:367–398
2. Benzley SE, Harris NJ, Scott M, Borden M, Owen SJ (2005) Conformal refinement and coarsening of unstructured hexahedral meshes. J Comput Inf Sci Eng 5:330–337
3. Bey J (1995) Tetrahedral grid refinement. Computing 55:355–378
4. Bielak J, Loukakis K, Hisada Y, Yoshimura C (2003) Domain reduction method for three-dimensional earthquake modeling in localized regions, Part I: Theory. Bull Seism Soc Am 93:817–824

5. Bonilla LF, Archuleta RJ, Lavallée D (2005) Hysteretic and dilatant behavior of cohesionless soils and their effects on nonlinear site response: Field data observations and modelling. Bull Seism Soc Am 95(6):2373–2395
6. Boore D (1972) Finite-difference methods for seismic wave propagation in heterogeneous materials. In: Bolt BA (ed) Methods in Computational Physics, vol 11. Academic Press, New York
7. Braun J, Sambridge MS (1995) A numerical method for solving partial differential equations on highly irregular evolving grids. Nature 376:655–660
8. Bunge HP, Tromp J (2003) Supercomputing moves to universities and makes possible new ways to organize computational research. EOS 84(4):30, 33
9. Carcione JM, Wang J-P (1993) A Chebyshev collocation method for the elastodynamic equation in generalised coordinates. Comp Fluid Dyn 2:269–290
10. Carcione JM, Kosloff D, Kosloff R (1988) Viscoacoustic wave propagation simulation in the earth. Geophysics 53: 769–777
11. Carcione JM, Kosloff D, Behle A, Seriani G (1992) A spectral scheme for wave propagation simulation in 3-D elastic-anisotropic media. Geophysics 57:1593–1607
12. Carey G (1997) Computational grids: Generation, adaptation, and solution strategies. Taylor Francis, New York
13. Cerveny V (2001) Seismic ray theory. Cambridge University Press, Cambridge
14. Chaljub E, Tarantola A (1997) Sensitivity of SS precursors to topography on the upper-mantle 660-km discontinuity. Geophys Res Lett 24(21):2613–2616
15. Chaljub E, Komatitsch D, Vilotte JP, Capdeville Y, Valette B, Festa G (2007) Spectral element analysis in seismology. In: Wu R-S, Maupin V (eds) Advances in wave propagation in heterogeneous media. Advances in Geophysics, vol 48. Elsevier, London, pp 365–419
16. Chapman CH (2004) Fundamentals of seismic wave propagation. Cambridge University Press, Cambridge
17. CIG www.geodynamics.org. Accessed 1 Jul 2008
18. Cockburn B, Shu CW (1989) TVB Runge"Kutta local projection discontinuous Galerkin finite element method for conservation laws II: General framework. Math Comp 52:411–435
19. Cockburn B, Shu CW (1991) The Runge–Kutta local projection P1-Discontinuous Galerkin finite element method for scalar conservation laws. Math Model Numer Anal 25:337–361
20. Cockburn B, Shu CW (1998) The Runge–Kutta discontinuous Galerkin method for conservation laws V: Multidimensional systems. J Comput Phys 141:199–224
21. Cockburn B, Lin SY, Shu CW (1989) TVB Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws III: One dimensional systems. J Comput Phys 84:90–113
22. Cockburn B, Hou S, Shu CW (1990) The Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws IV: The multidimensional case. Math Comp 54:545–581
23. Cockburn B, Karniadakis GE, Shu CW (2000) Discontinuous Galerkin methods, theory, computation and applications. LNCSE, vol 11. Springer, New York
24. Courant R, Friedrichs KO, Lewy H (1928) Über die partiellen Differenzialgleichungen der mathematischen Physik. Mathematische Annalen 100:32–74
25. CUBIT cubit.sandia.gov. Accessed 1 Jul 2008
26. Dablain MA (1986) The application of high-order differencing to the scalar wave equation. Geophysics 51:54–66
27. De Cougny HL, Shephard MS (1999) Parallel refinement and coarsening of tetrahedral meshes. Int J Numer Methods Eng 46:1101–1125
28. de la Puente J, Dumbser M, Käser M, Igel H (2007) Discontinuous Galerkin methods for wave propagation in poroelastic media. to appear in Geophysics
29. de la Puente J, Käser M, Dumbser M, Igel H (2007) An arbitrary high order discontinuous Galerkin method for elastic waves on unstructured meshes IV: Anisotropy. Geophys J Int 169(3):1210–1228
30. di Prisco C, Stupazzini M, Zambelli C (2007) Non-linear SEM numerical analyses of dry dense sand specimens under rapid and dynamic loading. Int J Numer Anal Methods Geomech 31(6):757–788
31. Dormy E, Tarantola A (1995) Numerical simulation of elastic wave propagation using a finite volume method. J Geophys Res 100(B2):2123–2134
32. Dumbser M (2005) Arbitrary high order schemes for the solution of hyperbolic conservation laws in complex domains. Shaker, Aachen
33. Dumbser M, Käser M (2006) An arbitrary high order discontinuous galerkin method for elastic waves on unstructured meshes II: The three-dimensional isotropic case. Geophys J Int 167:319–336
34. Dumbser M, Käser M (2007) Arbitrary high order non-oscillatory finite volume schemes on unstructured meshes for linear hyperbolic systems. J Comput Phys 221:693–723. doi:10.1016/j.jcp.2006.06.043
35. Dumbser M, Munz CD (2005) Arbitrary high order discontinuous Galerkin schemes. In: Cordier S, Goudon T, Gutnic M, Sonnendrucker E (eds) Numerical methods for hyperbolic and kinetic problems. IRMA series in mathematics and theoretical physics. EMS Publishing, Zurich, pp 295–333
36. Dumbser M, Käser M, Toro EF (2007) An arbitrary high-order discontinuous Galerkin method for elastic waves on unstructured meshes – V. Local time stepping and p-adaptivity. Geophys J Int 171:695–717
37. Dziewonski AM, Anderson DL (1981) Preliminary reference earth model. Phys Earth Planet Inter 25:297–356
38. Ewald M, Igel H, Hinzen K-G, Scherbaum F (2006) Basin-related effects on ground motion for earthquake scenarios in the lower rhine embayment. Geophys J Int 166:197–212
39. Faccioli E, Maggio F, Quarteroni A, Tagliani A (1996) Spectral-domain decomposition methods for the solution of acoustic and elastic wave equation. Geophysics 61:1160–1174
40. Faccioli E, Maggio F, Paolucci R, Quarteroni A (1997) 2D and 3D elastic wave propagation by a pseudo-spectral domain decomposition method. J Seismol 1:237–251
41. Faccioli E, Vanini M, Paolucci R, Stupazzini M (2005) Comment on "Domain reduction method for three-dimensional earthquake modeling in localized regions, part I: Theory." by Bielak J, Loukakis K, Hisada Y, Yoshimura C, and "Part II: Verification and Applications." by Yoshimura C, Bielak J, Hisada Y, Fernández A. Bull Seism Soc Am 95:763–769
42. Falk J, Tessmer E, Gajewski D (1996) Efficient finite-difference modelling of seismic waves using locally adjustable time steps. Geophys Prosp 46:603–616

43. Falk J, Tessmer E, Gajewski D (1996) Tube wave modelling by the finite differences method with varying grid spacing. Pure Appl Geoph 148:77–93
44. Fernandez A, Bielak J, Prentice C (2006) Urban seismology; City effects on earthquake ground motion and effects of spatial distribution of ground motion on structural response paper presented at 2006 annual meeting. Seism Res Lett 77(2):305
45. Fornberg B (1996) A practical guide to pseudospectral methods. Cambridge University Press, Cambridge
46. Fuchs K, Müller G (1971) Computation of synthetic seismograms with the reflectivity method and comparison with observations. Geophys J Royal Astronom Soc 23(4):417–33
47. Furumura T, Takenaka H (1996) 2.5-D modeling of elastic waves using the pseudospectral method. Geophys J Int 124: 820–832
48. Geller RJ, Takeuchi N (1998) Optimally accurate second-order time-domain finite difference scheme for the elastic equation of motion: One-dimensional case. Geophys J Int 135:48–62
49. Graves RW (1993) Modeling three-dimensional site response effects in the Marina district basin, San Francisco, California. Bull Seism Soc Am 83:1042–1063
50. Hestholm SO, Ruud BO (1998) 3-D finite-difference elastic wave modeling including surface topography. Geophysics 63:613–622
51. Holberg O (1987) Computational aspects of the choice of operator and sampling interval for numerical differentiation in large-scale simulation of wave phenomena. Geophys Prospect 35:629–655
52. Igel H (1999) Wave propagation through 3-D spherical sections using the Chebyshev spectral method. Geop J Int 136:559–567
53. Igel H, Gudmundsson O (1997) Frequency-dependent effects on travel times and waveforms of long-period S and SS waves. Phys Earth Planet Inter 104:229–246
54. Igel H, Weber M (1995) SH-wave propagation in the whole mantle using high-order finite differences. Geophys Res Lett 22(6):731–734
55. Igel H, Weber M (1996) P-SV wave propagation in the Earth's mantle using finite-differences: Application to heterogeneous lowermost mantle structure. Geophys Res Lett 23:415–418
56. Igel H, Mora P, Riollet B (1995) Anisotropic wave propagation through finite-difference grids. Geophysics 60:1203–1216
57. Igel H, Nissen-Meyer T, Jahnke G (2001) Wave propagation in 3-D spherical sections: Effects of subduction zones. Phys Earth Planet Inter 132:219–234
58. Jahnke G, Igel H, Cochard A, Thorne M (2007) Parallel implementation of axisymmetric SH wave propagation in spherical geometry. Geophys J Int (in print)
59. Jastram C, Tessmer E (1994) Elastic modelling on a grid with vertically varying spacing. Geophys Prosp 42:357–370
60. Karypis G, Kumar V (1998) Multilevel k-way Partitioning Scheme for Irregular Graphs. J Parallel Distrib Comput 48(1):96–129
61. Kelly KR, Ward RW, Treitel S, Alford RM (1976) Synthetic seismograms: A finite-difference approach. Geophysics 41: 2–27
62. Kennett BLN (2002) The seismic wavefield, vol I + II. Cambridge University Press, Cambridge
63. Komatitsch D, Tromp J (2002) Spectral-element simulations of global seismic wave propagation, part I: Validation. Geophys J Int 149:390–412
64. Komatitsch D, Tromp J (2002) Spectral-element simulations of global seismic wave propagation, part II: 3-D models, oceans, rotation, and gravity. Geophys J Int 150:303–318
65. Komatitsch D, Vilotte JP (1998) The spectral-element method: An efficient tool to simulate the seismic response of 2D and 3D geological structures. Bull Seism Soc Am 88:368–392
66. Komatitsch D, Coutel F, Mora P (1996) Tensorial formulation of the wave equation for modelling curved interfaces. Geophys J Int 127(1):156–168
67. Kosloff D, Baysal E (1982) Forward modeling by a fourier method. Geophysics 47(10):1402–1412
68. Krishnan S, Ji C, Komatitsch D, Tromp J (2006) Case studies of damage to tall steel moment-frame buildings in Southern California during large San Andreas earthquakes. Bull Seismol Soc Am 96(4A):1523–1537
69. Krishnan S, Ji C, Komatitsch D, Tromp J (2006) Performance of two 18-story steel moment-frame buildings in Southern California during two large simulated San Andreas earthquakes. Earthq Spectra 22(4):1035–106
70. Käser M, Dumbser M (2006) An arbitrary high order discontinuous Galerkin method for elastic waves on unstructured meshes I: The two-dimensional isotropic case with external source terms. Geophys J Int 166:855–877
71. Käser M, Dumbser M, de la Puente J, Igel H (2007) An arbitrary high order discontinuous Galerkin method for elastic waves on unstructured meshes III: Viscoelastic attenuation. Geophys J Int 168(1):224–242
72. Käser M, Igel H (2001) Numerical simulation of 2D wave propagation on unstructured grids using explicit differential operators. Geophys Prospect 49(5):607–619
73. Käser M, Igel H, Sambridge M, Braun J (2001) A comparative study of explicit differential operators on arbitrary grids. J Comput Acoust 9(3):1111–1125
74. Kwak D-Y, Im Y-T (2002) Remeshing for metal forming simulations – part II: Three dimensional hexahedral mesh generation. Int J Numer Methods Eng 53:2501–2528
75. LeVeque RL (2002) Finite volume methods for hyperbolic problems. Cambridge University Press, Cambridge
76. Levander AR (1988) Fourth-order finite-difference P-SV seismograms. Geophysics 53:1425–1436
77. Madariaga R (1976) Dynamics of an expanding circular fault. Bull Seismol Soc Am 66(3):639–66
78. Magnier S-A, Mora P, Tarantola A (1994) Finite differences on minimal grids. Geophysics 59:1435–1443
79. Marfurt KJ (1984) Accuracy of finite-difference and finite-element modeling of the scalar and elastic wave equations. Geophysics 49:533–549
80. Mercerat ED, Vilotte JP, Sanchez-Sesma FJ (2006) Triangular spectral element simulation of two-dimensional elastic wave propagation using unstructured triangular grids. Geophys J Int 166(2):679–698
81. METIS glaros.dtc.umn.edu/gkhome/views/metis. Accessed 1 Jul 2008
82. Moczo P (1989) Finite-difference techniques for SH-waves in 2-D media using irregular grids – Application to the seismic response problem. Geophys J Int 99:321–329

83. Moczo P, Kristek J, Halada L (2000) 3D 4th-order staggered grid finite-difference schemes: Stability and grid dispersion. Bull Seism Soc Am 90:587–603
84. Montelli R, Nolet G, Dahlen FA, Masters G, Engdahl ER, Hung S (2004) Finite-frequency tomography reveals a variety of plumes in the mantle. Science 303(5656):338–343
85. Müller G (1977) Earth-flattening approximation for body waves derived from geometric ray theory – improvements, corrections and range of applicability. J Geophys 42:429–436
86. Nissen-Meyer T, Fournier A, Dahlen FA (2007) A 2-D spectral-element method for computing spherical-earth seismograms – I. Moment-tensor source. Geophys J Int 168:1067–1092
87. Ohminato T, Chouet BA (1997) A free-surface boundary condition for including 3D topography in the finite-difference method. Bull Seism Soc Am 87:494–515
88. Opršal I, J Zahradník (1999) Elastic finite-difference method for irregular grids. Geophysics 64:240–250
89. Pitarka A (1999) 3D elastic finite-difference modeling of seismic motion using staggered grids with nonuniform spacing. Bull Seism Soc Am 89:54–68
90. Priolo E, Carcione JM, Seriani G (1996) Numerical simulation of interface waves by high-order spectral modeling techniques. J Acoust Soc Am 95:681–693
91. Reed WH, Hill TR (1973) Triangular mesh methods for the neutron transport equation. Technical Report, LA-UR-73-479, Los Alamos Scientific Laboratory
92. Ripperger J, Igel H, Wassermann J (2004) Seismic wave simulation in the presence of real volcano topography. J Volcanol Geotherm Res 128:31–44
93. Scandella L (2007) Numerical evaluation of transient ground strains for the seismic response analyses of underground structures. Ph D Thesis, Milan University of Technology, Milan
94. SCEC www.scec.org. Accessed 1 Jul 2008
95. Schneiders R (2000) Octree-Based Hexahedral Mesh Generation. Int J Comput Geom Appl 10(4):383–398
96. Schwartzkopff T, Munz CD, Toro EF (2002) ADER: A high-order approach for linear hyperbolic systems in 2D. J Sci Comput 17:231–240
97. Schwartzkopff T, Dumbser M, Munz CD (2004) Fast high order ADER schemes for linear hyperbolic equations. J Comput Phys 197:532–539
98. Seriani G, Priolo E, Carcione JM, Padovani E (1992) High-order spectral element method for elastic wave modeling: 62nd Ann. Internat. Mtg., Soc. Expl. Geophys., Expanded Abstracts, 1285–1288
99. Shepherd JF (2007) Topologic and geometric constraint-based hexahedral mesh generation. Ph.D. Thesis on Computer Science, School of Computing The Universiy of Utah, Salt Lake City
100. Sieminski A, Liu Q, Trampert J, Tromp J (2007) Finite-frequency sensitivity of surface waves to anisotropy based upon adjoint methods. Geophys J Int 168:1153–1174
101. SPICE www.spice-rtn.org. Accessed 1 Jul 2008
102. Stupazzini M (2004) A spectral element approach for 3D dynamic soil-structure interaction problems. Ph D Thesis, Milan University of Technology, Milan
103. Takeuchi N, Geller RJ (2000) Optimally accurate second order time-domain finite difference scheme for computing synthetic seismograms in 2-D and 3-D media. Phys Earth Planet Int 119:99–131
104. Tape C, Liu Q, Tromp J (2007) Finite-frequency tomography using adjoint methods: Methodology and examples using membrane surface waves. Geophys J Int 168:1105–1129
105. Tarantola A (1986) A strategy for nonlinear elastic inversion of seismic reflection data. Geophysics 51(10):1893–1903
106. Tessmer E (2000) Seismic finite-difference modeling with spatially varying time teps. Geophysics 65:1290–1293
107. Tessmer K, Kosloff D (1996) 3-D elastic modeling with surface topography by a Chebychev spectral method. Geophysics 59:464–473
108. Tessmer E, Kessler D, Kosloff K, Behle A (1996) Multi-domain Chebyshev–Fourier method for the solution of the equations of motion of dynamic elasticity. J Comput Phys 100:355–363
109. Thomas C, Igel H, Weber M, Scherbaum F (2000) Acoustic simulation of P-wave propagation in a heterogeneous spherical earth: Numerical method and application to precursor energy to PKPdf. Geophys J Int 141:307–320
110. Thorne M, Lay T, Garnero E, Jahnke G, Igel H (2007) 3-D seismic imaging of the D″ region beneath the Cocos Plate. Geophys J Int 170:635–648
111. Titarev VA, Toro EF (2002) ADER: Arbitrary high order Godunov approach. J Sci Comput 17:609–618
112. Toro EF (1999) Riemann solvers and numerical methods for fluid dynamics. Springer, Berlin
113. Toro EF, Millington AC, Nejad LA (2001) Towards very high order Godunov schemes, in Godunov methods; Theory and applications. Kluwer/Plenum, Oxford, pp 907–940
114. Toyokuni G, Takenaka H, Wang Y, Kennett BLN (2005) Quasi-spherical approach for seismic wave modeling in a 2-D slice of a global earth model with lateral heterogeneity. Geophys Res Lett 32:L09305
115. Van der Hilst RD (2004) Changing views on Earth's deep mantle. Science 306:817–818
116. Virieux J (1984) SH-wave propagation in heterogeneous media: Velocity-stress inite-difference method. Geophysics 49:1933–1957
117. Virieux J (1986) P-SV wave propagation in heterogeneous media: Velocity-stress finite-difference method. Geophysics 51:889–901
118. Woodhouse JH, Dziewonski AM (1984) Mapping the upper mantle: Three dimensional modelling of earth structure by inversion of seismic waveforms. J Geophys Res 89:5953–5986
119. Yoshimura C, Bielak J, Hisada Y, Fernández A (2003) Domain reduction method for three-dimensional earthquake modeling in localized regions, part II: Verification and applications. Bull Seism Soc Am 93:825–841
120. Zambelli C (2006) Experimental and theoretical analysis of the mechanical behaviour of cohesionless soils under cyclic-dynamic loading. Ph D Thesis, Milan University of Technology, Milan
121. Zienckiewicz O, Taylor RL (1989) The finite element method, vol 1. McGraw-Hill, London

Books and Reviews

Carcione JM, Herman GC, ten Kroode APE (2002) Seismic modelling. Geophysics 67:1304–1325

Mozco P, Kristek J, Halada L (2004) The finite-difference method for seismologists: An introduction. Comenius University, Bratislava. Available in pdf format at ftp://ftp.nuquake.eu/pub/Papers

Moczo P, Kristek J, Galis M, Pazak P, Balazovjech M (2007) The finite difference and finite-element modelling of seismic wave propagation and earthquake motion. Acta Physica Slovaca, 57(2)177–406

Wu RS, Maupin V (eds) (2006) Advances in wave propagation in heterogeneous earth. In: Dmowska R (ed) Advances in geophysics, vol 48. Academic/Elsevier, London

Seismic Waves in Heterogeneous Earth, Scattering of

Haruo Sato
Department of Geophysics, Graduate School of Science,
Tohoku University, Sendai-shi, Miyagi-ken, Japan

Article Outline

Glossary

Attenuation factor Q^{-1} A measure of attenuation characteristics of a medium caused by intrinsic absorption and scattering loss. The former means the transfer of vibration energy into heat and the latter means the transfer of vibration energy from the direct wave to coda waves caused by scattering due to medium heterogeneity.

Coda waves Wave trains that follow the arrival of the direct S-wave phase are called S-coda waves or simply coda waves. Coda waves are interpreted as a superposition of S waves scattered by distributed heterogeneities. Wave trains between direct P and S wave arrivals are called P-coda waves.

Coda attenuation factor Q_C^{-1} This parameter characterizes the amplitude decay of S coda of a local earthquake with the lapse time increasing based on the S-to-S single scattering. The coda duration shortens for a larger coda attenuation factor.

Envelope broadening The source duration time of a microearthquake is short; however, the apparent duration time of the S-wave seismogram increases with the travel distance increasing because of diffraction and scattering by medium heterogeneities. This phenomenon is called envelope broadening.

Radiative transfer theory A phenomenological theory that describes scattering process of wave energy in a scattering medium on the basis of causality, geometrical spreading and the energy conservation. It neglects the interference of waves but focuses on the intensity only. This theory admits various types of scattering patterns. It is often applied to model the energy propagation of high-frequency seismic-waves in heterogeneous Earth media.

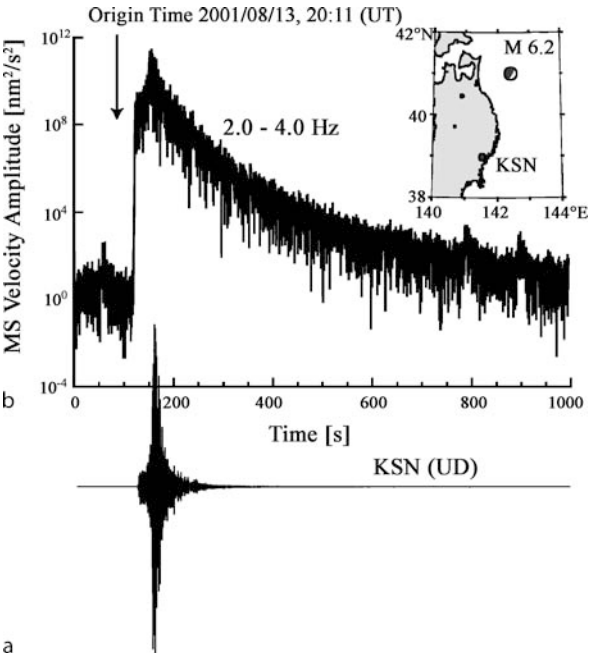
Random media A mathematical model for media whose parameters are described by random functions of space coordinates. The stochastic properties of the ensemble of random media are characterized by their autocorrelation function or the power spectral density function.

Scattering coefficient g A measure of the scattering power in a unit solid angle at a certain direction by a unit volume of heterogeneous media for the incidence of unit energy flux density. The average of g over the solid angle gives the total scattering coefficient g_0, of which the reciprocal gives the mean free path. This quantity characterizes the coda excitation and the scattering loss in the heterogeneous media.

Definition of the Subject

The structure of the solid Earth was extensively studied by using seismic waves such as travel time analysis based on Snell's law, dispersion analysis of surface waves, and spectral analysis of free oscillation, where the notion of a horizontally stratified structure or a spherical shell structure prevailed among the geophysical community. This means the acceptance of the dominance of gravity in geodynamic process. Velocity tomography revealed that the solid Earth structure is three-dimensionally inhomogeneous with various ranges of scales; however, the resolution of velocity tomography is much coarser than the wavelength of seismic waves. In 1970s, the existence of distributed inhomogeneities having the order of the wavelength of seismic waves was recognized from the observation of coda waves of local earthquakes, which are long-lasting wave trains following the direct S-wave arrival in high-frequency seismograms. Here, we use "high-frequency" for frequency higher than about 1 Hz. The long duration time of coda waves can be interpreted as a direct evidence of wide-angle scattering caused by distributed small-scale heterogeneities since the source duration time is generally very short. S-wave seismograms of microearthquakes show broadened envelopes with travel distance increasing. This envelope broadening phenomenon is also an evidence of scattering around the forward direction due to random velocity inhomogeneities.

Since then, focusing on the frequency dependence of seismogram envelopes, geophysicists have extensively studied the scattering process of high-frequency seismic waves in relation to the spectral structure of velocity inhomogeneities, where the statistical characterization of the medium inhomogeneity is inevitable. The radiative transfer theory and the stochastic Markov approximation have been developed as mathematical tools for the analyzes of seismogram envelopes. The strength of scattering and/or the spectral structure of random inhomogeneity have been measured in various regions of the solid Earth. The scattering approach is found to be also useful for detecting temporal changes in the crustal medium associated with earthquake occurrences. Thus, scattering of high-frequency seismic waves in the heterogeneous Earth medium is important for understanding the physical structure and the geodynamic process reflecting the evolution of the solid Earth.



Seismic Waves in Heterogeneous Earth, Scattering of, Figure 1
a Seismogram of a local earthquake of M 6.2 in northeastern Honshu, Japan recorded by F-net, NIED. b Bandpass-filtered MS envelope (Courtesy of T. Maeda)

Introduction

Coda Waves

The high-frequency seismogram of a local earthquake has a long tail after the direct S-coda arrival. The tail portion of seismogram is called “S-coda waves” or simply “coda”. As an example, Fig. 1a and b show the raw seismogram and the band-pass filtered mean square (MS) trace of an earthquake of magnitude (M) 6.1, respectively. We note that the mean square wave envelope, which is the running mean of the squared trace with characteristic time of a few times the center period, is proportional to the time trace of the wave energy density. The coda wave oscillation lasts more than several hundreds of seconds. The duration of coda waves measured from the P-wave onset until when the coda amplitude decreases to the microseism’s level has been used as a quick measure of the earthquake magnitude from a single station observation since the 1960s. Having a motivation to extract the source spectrum of a large earthquake from clipped seismograms, Aki [1] first studied the characteristics of coda waves as scattered waves. Coda envelopes of a local earthquake have a smoothly decaying common curve with lapse time increasing irrespective of epicentral distances and the source radiation pattern. Aki and Chouet [4] interpreted coda waves as single back-scattered S-waves due to heterogeneities randomly distributed in the lithosphere. Their model based on the radar equation for the same location of a source and a receiver can be written as follows.

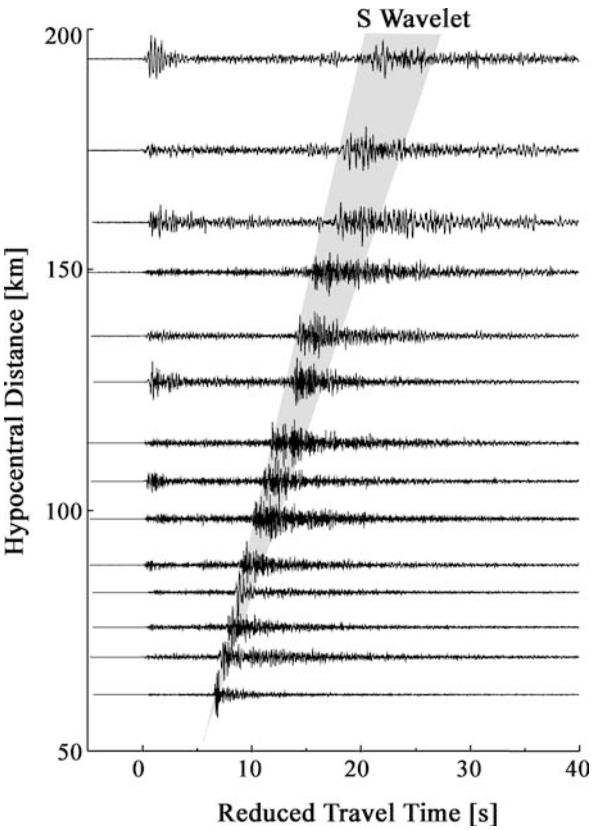
Point-like isotropic scatterers characterized by total scattering cross-section σ_0 are randomly and homogeneously distributed with number density n in a 3-D medium with background wave velocity V_0. The scattering power per unit volume is characterized by the total scattering coefficient $g_0 = n\sigma_0$, of which the reciprocal gives the mean free path. When the total wave energy W is impulsively radiated from a point source at time $t = 0$, the wave energy density of singly back-scattered waves at the source location is written as

$$E^{\mathrm{SB}}(t) \approx \frac{W g_0}{2\pi V_0^2 t^2} \mathrm{e}^{-Q_{\mathrm{C}}^{-1} 2\pi f t} \tag{1}$$

since the interference of scattered waves can be neglected because of the random distribution of scatterers. The inverse square of lapse time means geometrical spreading in a 3-D space. Here, an exponential damping term with coda attenuation factor Q_{C}^{-1} is introduced to represent phenomenological attenuation effect. This simple formula has been widely used for measurements of g_0 and Q_{C}^{-1} for S-waves in the world since the 1980s. Reported g_0-values are of the order of 0.01 km^{-1} for 1–30 Hz and Q_C^{-1} values are about 10^{-2} at 1 Hz and decrease to about 10^{-3} at 20 Hz in the lithosphere [74].

Envelope Broadening of S-Seismogram

There is another evidence of scattering due to random inhomogeneity in high-frequency seismograms. Observed S-seismograms of a microearthquake have broadened envelopes around their peaks after the direct arrivals. As shown by an example in Fig. 2, the apparent duration time of the S-seismogram just after the direct S-arrival increases with increasing travel distance. It is more than ten seconds at distances larger than 100 km, where the source duration time is less than one second for an earthquake of M 4.0. Sato [70] called this phenomenon observed in an island arc as "envelope broadening", and Atkinson [6] reported similar phenomenon in a continent. For P-waves of teleseismic events, broadening of the vertical component envelope [35] and the excitation in the transverse component amplitude [52] have been used as a measure of lithospheric heterogeneity. These phenomena can be interpreted by multiple scattering within a narrow angle around the global ray direction due to random velocity inhomogeneities. When the wavelength is much shorter than the characteristic scale of the random velocity inhomogeneity, the scattering process of waves can be represented by successive ray bending processes, where scattering angles are statistically controlled by the spectrum of random velocity inhomogeneity. At a given distance from the source, a small number of rays with large scattering angles arrive long after the direct ray.



Seismic Waves in Heterogeneous Earth, Scattering of, Figure 2
Envelope broadening shown in horizontal component seismograms of a microearthquake with M 4.0 in Japan recorded by Hi-net, NIED, where the abscissa is reduced travel time with move-out velocity 7 km/s (Courtesy of T. Takahashi)

Radiative Transfer Theory for a Scattering Medium

Radiative Transfer Integral Equation for the Isotropic Scattering Process

Disregarding wave interference and focusing on wave power, the radiative transfer theory [8] treats the propagation of wave energy in a scattering medium. Wu [85] first introduced the radiative transfer theory for the stationary state in the synthesis of seismogram envelopes. The non-stationary multiple isotropic scattering process in 1-D was solved by Hemmer [24] and that in 2-D was solved by Shang and Gao [77]. Later, Zeng et al. [92] formulated the time-dependent multiple isotropic scattering process in 3-D as an extension of the single backscattering model [4] as follows.

In a 3-D scattering medium characterized by background velocity V_0 and total scattering coefficient g_0, when the total wave energy W is impulsively radiated isotropically from a source at the origin, the multiple isotropic scattering process is written by the following integral equation for energy density:

$$E(\mathbf{x},t) = WG_E(\mathbf{x},t) + g_0 V_0 \cdot \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G_E\left(\mathbf{x}-\mathbf{x}', t-t'\right) E\left(\mathbf{x}', t'\right) \mathrm{d}\mathbf{x}'\mathrm{d}t' , \tag{2}$$

where the convolution integral in the second term means the propagation of energy from the last scattering point $\mathbf{x}'$ to a receiver at $\mathbf{x}$. The first term is the ballistic term that means the direct propagation of energy from the source with scattering loss, $G_E(\mathbf{x},t) = \delta(t - r/V_0)\exp(-g_0 V_0 t)/(4\pi V_0 r^2)$, where $r = |\mathbf{x}|$. The solution is written as [92]

$$
\begin{aligned}
E(\mathbf{x}, t) = {} & W G_{\mathrm{E}}(\mathbf{x}, t) + \frac{W g_0 \mathrm{e}^{-g_0 V_0 t}}{4\pi r^2}\frac{r}{V_0 t} \\
& \cdot \ln\left[\frac{V_0 t + r}{V_0 t - r}\right] H\left(t - \frac{r}{V_0}\right) \\
& + W g_0^2 V_0^2 \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{d}\omega \mathrm{d}k \, \mathrm{e}^{-i\omega t - ikr} \\
& \cdot \frac{ik}{2\pi r}\frac{\bar{\bar{G}}_{\mathrm{E}}(-k, -i\omega)^3}{1 - g_0 V_0 \bar{\bar{G}}_{\mathrm{E}}(-k, -i\omega)} ,
\end{aligned}
\tag{3}
$$

where $\bar{\bar{G}}_{\mathrm{E}}(k, s) = (1/k V_0)\tan^{-1} k V_0/\left(s + g_0 V_0\right)$ is the Fourier–Laplace transform of G_{E} with respect to coordinate and time, respectively. The second term represents the single scattering process (see Sato [67]), which has a logarithmic divergence at the direct arrival $t = r/V_0$ and decreases according to the inverse square of lapse time at long lapse times as $W g_0/\left(2\pi V_0^2 t^2\right)$. The third term representing multiple scattering converges to a diffusion solution with lapse time increasing as

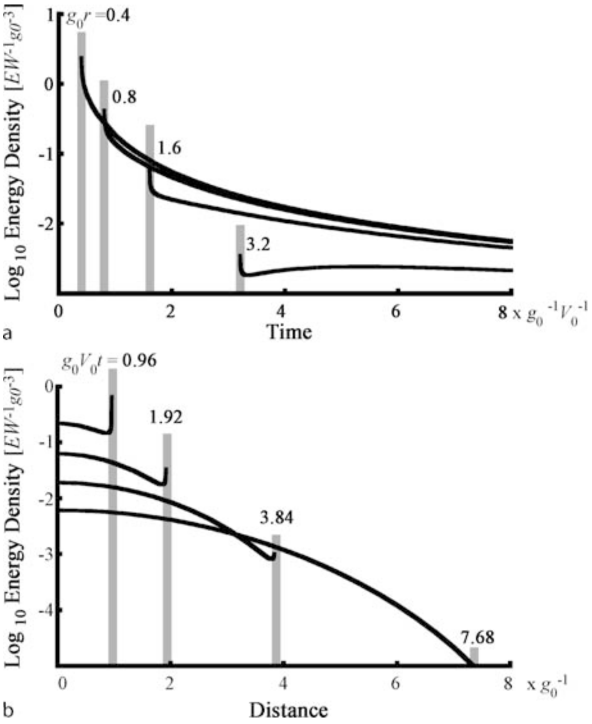
$$
E^{\mathrm{Dif.}}(\mathbf{x}, t) = W\left(\frac{3 g_0}{4\pi V_0 t}\right)^{3/2} \mathrm{e}^{-\frac{3 g_0 r^2}{4 V_0 t}} H(t) ,
\tag{4}
$$

where the factor $V_0/\left(3 g_0\right)$ is the diffusivity. Later, Paaschens [54] proposed an approximation as

$$
\begin{aligned}
E^{\mathrm{Paa.}}(\mathbf{x}, t) \approx {} & W G_{\mathrm{E}}(\mathbf{x}, t) + \frac{W \mathrm{e}^{-g_0 V_0 t}}{\left(4\pi V_0 t/\left(3 g_0\right)\right)^{3/2}} \\
& \cdot \left(1 - \frac{r^2}{V_0^2 t^2}\right)^{\frac{1}{8}} M\left(g_0 V_0 t\left(1 - \frac{r^2}{V_0^2 t^2}\right)^{\frac{3}{4}}\right) \\
& \cdot H\left(t - \frac{r}{V_0}\right) ,
\end{aligned}
\tag{5}
$$

where $M(x) \approx \mathrm{e}^x \sqrt{1 + 2.026/x}$. The error of this approximation is of the order of 2% outside the ballistic peak and its tail for $g_0 V_0 t < 6$ and $2 < g_0 r < 4$.

Figure 3 shows spatiotemporal variations in energy density in a scattering medium theoretically predicted by the approximation solution (5) for instantaneous spherical source radiation at the origin. Scattered energy density is shown by a black curve, where the ballistic term is shown by a vertical gray line. At a small distance from the source compared with the mean free path $1/g_0$, the energy density decreases rapidly after the direct arrival as predicted by the single scattering term; however, the decay rate becomes smaller due to multiple scattering according to the power of lapse time $t^{-3/2}$ at long lapse times. At a long distance, for example at $r = 3.2/g_0$, the energy density has an additional diffusion peak. The spatial distribution of scattered energy density is uniform around the source at a short lapse time compared with the mean free time $1/g_0 V_0$; however, it converges to a Gaussian curve at a long lapse time as theoretically predicted by the diffusion solution (4), for example at $t = 7.48/g_0 V_0$. There is no violation of causality since no signal exists beyond the ballistic peak. The smooth spatial distribution of scattered energy density around the source location gives the physical basis of the coda normalization method for measurements of S-wave attenuation and site amplification factors (e. g. [3,56,89]).



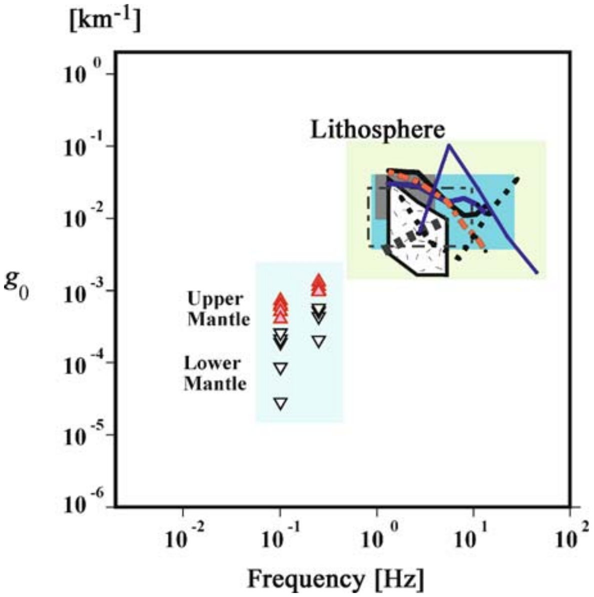
Seismic Waves in Heterogeneous Earth, Scattering of, Figure 3
a Temporal change and b spatial variation of energy density in an isotropic-scattering medium for a point source radiation. Each *black curve* shows the scattering contribution predicted by the Paaschens approximation and each *vertical gray line* shows a ballistic term

Gusev and Abubakirov [21] and Hoshiba [27] numerically solved the radiative transfer equation for the isotropic scattering process by using the Monte Carlo method. Yoshimoto [88] numerically simulated envelopes in scattering media of which the background velocity decreases with depth. He found the concentration of scattered energy near the surface because of seismic ray bending.

Nonisotropic radiation from a source can be easily introduced in the radiative transfer equation. Sato et al. [75] analytically solved the case of double couple source radiation: the energy density theoretically predicted faithfully reflects the source radiation pattern near the direct arrival; however, the azimuthal dependence diminishes with increasing lapse time. It qualitatively agrees with the observed radiation pattern independence of coda amplitudes at long lapse times. Their solution has been used in the envelope inversion of strong motion records for the spatial distribution of high-frequency wave energy radiation from an earthquake fault (e. g. [47]).

Measurements of Total Scattering Coefficient and Attenuation

For the practical application of the radiative transfer theory to observed seismograms, it is necessary to introduce intrinsic absorption Q_{Int}^{-1} by multiplying an exponential temporal decay factor $\exp[-Q_{\mathrm{Int}}^{-1} 2\pi f t]$ to the resultant energy density. By using the solution (3) of the radiative transfer theory for the isotropic scattering model, total scattering coefficient g_0 and intrinsic absorption factor $Q_{\mathrm{Int.}}^{-1}$ of the S-wave have been measured. Reported g_0 values in the lithosphere are of the order of $10^{-2}\,\mathrm{km}^{-1}$ for frequencies from 1 to 30 Hz as plotted in Fig. 4.



Seismic Waves in Heterogeneous Earth, Scattering of, Figure 4
Total scattering coefficient of S-waves in the Earth. Measurements in the mantle [40] are added to lithospheric inhomogeneity [74]

From the observed lapse time dependence of Q_{C}^{-1}, Gusev [20] quantitatively explained the decrease of g_0 with depth. Lee et al. [40] analyzed coda envelopes of regional earthquakes before and after the ScS arrival around 900s in lapse time from the origin time based on the numerically simulated envelopes for the PREM model, which is characterized by depth-dependent background velocity and total attenuation for S-waves. They reported lower g_0 values in 4s and 10s period bands in the upper and lower mantle compared with those in the lithosphere as illustrated in Fig. 4.

Hoshiba et al. [28] and Fehler et al. [10] developed a method to measure simultaneously g_0 and Q_{Int}^{-1} values for the S-wave from the whole S-envelope analysis based on the synthetic envelope derived from the radiative transfer theory. Their multiple lapse-time window analysis method has been widely used in the world. Estimated scattering loss $g_0 V_0/\omega$ decreases with frequency; however, intrinsic absorption Q_{Int}^{-1} is rather insensitive to frequency. Estimated seismic albedo $B \equiv g_0 V_0 / \left(\omega Q_{\mathrm{Int}}^{-1} + g_0 V_0\right)$, the ratio of scattering loss to the total attenuation, of S-waves in the lithosphere widely distribute from 0.2 to 0.8 for 1–6 Hz, but they are limited between 0.2 and 0.5 for 6–20 Hz.

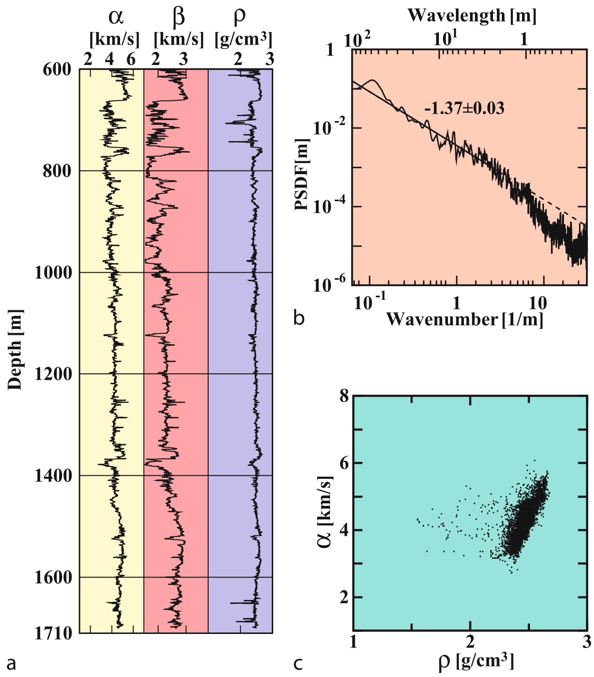
Total scattering cross-section and number density of scatterers appear jointly as the total scattering coefficient in theoretical models; however, Matsumoto [44] proposed a method to separate them from the temporal variation of the semblance coefficient of coda waves recorded by a seismic array. Analyzing data obtained in the aftershock area of the 2000 western Tottori earthquake, Japan, he estimated $n = 0.03\,\mathrm{km}^{-3}$ and $g_0 = 0.001\,\mathrm{km}^{-1}$ at 20 Hz.

It should be noted that lunar seismograms have coda durations exceeding one hour. Applying the diffusion model for the explanation of spindle-like envelopes of lunar-quakes, Dainty and Toksöz [9] estimated g_0 to be as large as 0.05–0.5 km^{-1} at 0.45 Hz.

Wave Envelopes in Random Media and Statistical Characterization

Statistical Characterization of Random Media

As revealed from tomography analyses (e. g. [93]), velocity structure is three-dimensionally inhomogeneous especially in the lithosphere. Well log data show typical samples of the shallow crust. Figure 5a shows well log data of P- and S-wave velocities and mass density obtained in Kyushu, Japan [80], where wave velocities and mass density are measured in a borehole by using ultrasonic waves and gamma rays, respectively. These well log data clearly show random fluctuation with short wavelengths.



Seismic Waves in Heterogeneous Earth, Scattering of, Figure 5

a Well log data at the YT-2 site, Kyushu, Japan. b Power spectral density function of the fractional fluctuation of P-wave velocity log. c Scattergram of P-wave velocity and mass density. Reproduced from [80]

As a natural consequence, we imagine random inhomogeneities widely distributed in the solid Earth medium.

By using computer power, wave propagation in random media has been numerically studied extensively. Using finite difference simulations of waves in random media, Frankel and Clayton [14] first examined the relation between coda excitation and scattering loss and the spectrum of random inhomogeneity. On the basis of numerical simulations, Frankel and Wennerberg [15] proposed the energy flux model that has a uniform distribution of scattered energy behind the direct waves for the analysis of high-frequency seismogram envelopes. Using a boundary integral method for the simulation of waves in a medium containing many cavities, Yomogida and Benites [87] examined a relation between coda attenuation and the distribution of cavities.

There is an alternative approach to treat statistically randomly inhomogeneous media. The wave-velocity is written as $V(\mathbf{x}) = V_0\{1 + \xi(\mathbf{x})\}$, where V_0 is the average velocity and fractional fluctuation $\xi(\mathbf{x})$ is a homogeneous and isotropic random function of space coordinate $\mathbf{x}$. We imagine an ensemble of random media $\{\xi\}$, which is statistically characterized by the autocorrelation function (ACF) $R(\mathbf{x}) \equiv \left\langle \xi\left(\mathbf{x}+\mathbf{x}'\right)\xi\left(\mathbf{x}'\right)\right\rangle$, where angular brackets mean the ensemble average. The MS fractional fluctuation $\varepsilon^2 \equiv R(0)$ and the correlation distance a are key parameters. The Fourier transform of ACF gives the power spectral density function (PSDF) P. The PSDF of the P-wave velocity fractional fluctuation of well log data shows a power-law characteristic at large wavenumbers as illustrated in Fig. 5b. P-wave velocity and mass density show a good correlation as shown in Fig. 5c. The statistical view is useful for representing geological data, too (e. g. [18,26]).

Scattering Coefficient Based on the Born Approximation

When the medium inhomogeneity is small $|\xi| \ll 1$, scalar wave ϕ is governed by the following wave equation:

$$\left(\Delta - \frac{1}{V_0^2}\partial_t^2\right)\phi + \frac{2}{V_0^2}\xi(\mathbf{x})\partial_t^2\phi = 0\,. \tag{6}$$

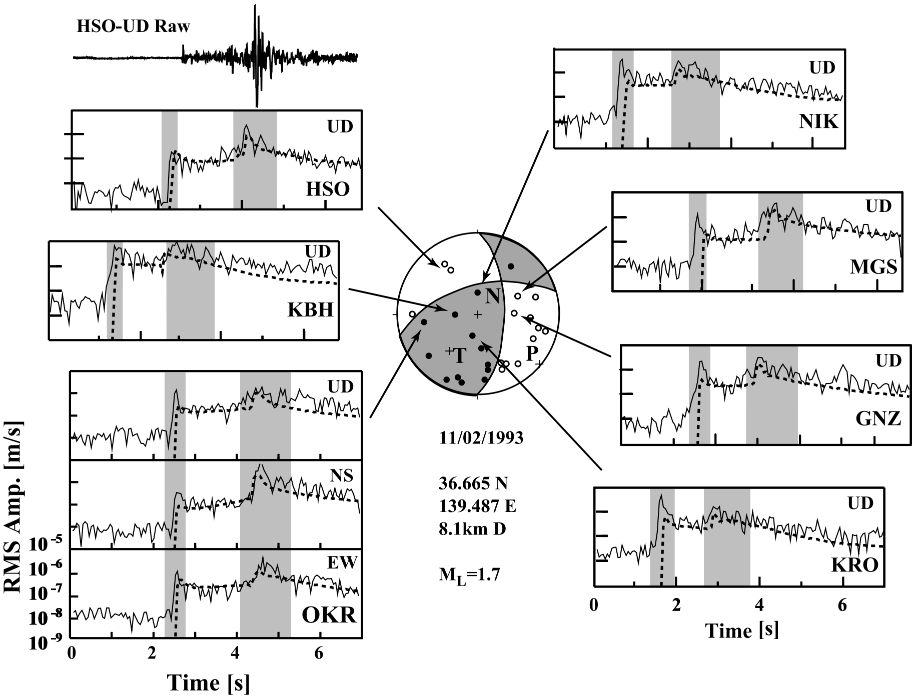
Velocity inhomogeneity is supposed to localize in a volume around the origin, of which the dimension is chosen to be much larger than a. For the incidence of a plane wave of unit amplitude at angular frequency ω as $e^{i(k_0\mathbf{e_z}\mathbf{x}-\omega t)}$, where $\mathbf{e_z}$ is the unit vector to the z direction, we calculate the spherically outgoing scattered waves due to a localized inhomogeneity by using the Born approximation as $\phi^1(\mathbf{x}, t) = -k_0^2 e^{i(k_0 r-\omega t)}\tilde{\xi}\left(k_0\mathbf{e_r} - k_0\mathbf{e_z}\right)/(2\pi r)$, where the tilde means the Fourier transform with respect to coordinates in 3-D space and $\mathbf{e_r}$ is a radial unit vector (e. g. [74]). According to Aki and Chouet [4], the scattering coefficient defined as the scattering power in a unit solid angle at certain direction by a unit volume of random inhomogeneous media for the incidence of unit energy flux density is statistically written by using its PSDF as

$$g(\psi;\omega) = \frac{k_0^4}{\pi}P\left(k_0\mathbf{e_r} - k_0\mathbf{e}_3\right) = \frac{k_0^4}{\pi}P\left(2k_0\sin\frac{\psi}{2}\right)\,, \tag{7}$$

where ψ is the scattering angle measured from the z direction. This functional form means anisotropic scattering depending on frequency. In general, scattering near around the forward direction becomes larger with increasing wavenumber in random media.

Radiative Transfer Theory with Scattering Coefficients Calculated by the Born Approximation

Extending the above formulation to vector wave propagation in random elastic media, we can define scattering coefficients for different scattering modes as PP, PS, SP,



Seismic Waves in Heterogeneous Earth, Scattering of, Figure 6
RMS envelopes of a microearthquake in the shallow crust, Nikko in northern Kanto, Japan. *Fine curves* and *broken curves* are observed and best-fit synthesized envelopes, respectively, where *shades* show time windows used for the estimation of the source radiation energy. The trace on the *top left* shows a raw seismogram. Reproduced from [90]

and SS. For the case of random elastic media characterized by an exponential ACF with $\varepsilon = 10\%$ and $a = 2$ km, Sato [68] synthesized three-component seismogram envelopes of a microearthquake of M 3 as a superposition of polarized scattered waves' power at a finite distance from a point shear dislocation source. That is the single scattering approximation of the radiative transfer theory with scattering coefficients calculated by the Born approximation. The SS scattering mode dominates in S coda, and pseudo P and S waves are produced even at a receiver on the null direction of the source radiation. By using a von Kármán-type ACF for describing random elastic media, Sato [71] estimated parameters $\kappa = 0.35, \varepsilon = 8.4\%$ and $a = 2.1$ km from observed frequency dependence of S-wave attenuation and g_0, where the parameter κ controls the role-off of PSDF at large wavenumbers. Extending the above vector wave envelope synthesis to include mode conversions at the free surface, Yoshimoto et al. [90] analyzed three-component seismogram envelopes of microearthquakes in the shallow crust in Nikko, northern Kanto, Japan. A raw seismogram is shown at the top-left of Fig. 6 as an example. Fine curves in Fig. 6 are logarithmic plots of observed root mean square (RMS) envelopes in the 2–16 Hz band. Broken curves are best-fit theoretical envelopes for random elastic media characterized by an exponential ACF with $\varepsilon = 5.7\%$ and $a = 400$ m. We find that the fitness is good not only for S coda but also for P coda.

Wave theory in random media predicts that the scattering coefficient has a large lobe in the forward direction in higher frequencies. Gusev and Abubakirov [22] used the Monte Carlo method to simulate envelopes for the multiple nonisotropic scattering process. There have been mathematical developments to derive the radiative transfer equation for multiple nonisotropic scattering from the stochastic averaging of the wave equation in random media (e. g. [13,29,41,64]). Przybilla et al. [58] showed an excellent coincidence of vector envelopes calculated from finite difference simulation in 2-D random elastic media and those synthesized by the radiative transfer theory with scattering amplitudes derived from the Born approximation and the wandering effect of travel time.

Interference of Scattered Waves

The interference effect is neglected in conventional studies of wave scattering in random media; however, it becomes important for a specific case even in random media. When randomness is strong enough to produce multiple scattering, coda wave intensity at a receiver near a source is enhanced compared to the prediction of conventional radiative transfer theory. Margerin et al. [42] showed that a spot of backscattering enhancement stabilizes in a sphere of radius half a wavelength centered at the source after a transient regime. The enhancement persists in time and should be observable as long as a coda is measurable. From field experiment of seismic waves in a shallow volcanic structure Larose et al. [39] reported the existence of weak localization, where the size of enhancement spot was one wavelength and the estimated mean-free path was 200 m for seismic waves around 20 Hz.

Envelope Broadening of a High-Frequency Seismogram

Markov Approximation for Parabolic Wave Equation

For the study of light propagation through the upper atmosphere and/or acoustic sound propagation through internal waves in oceans, various stochastic methods have been developed in the fields of physics. One of the most attractive methods for explaining the wave envelope around the direct arrival is the Markov approximation for the parabolic wave equation, which is an extension of the phase screen method or the split step Fourier method (e. g. [29,63]). This method is found to be applicable to seismogram envelopes. We imagine an elastic medium composed of a homogeneous half space $z < 0$ and an inhomogeneous half space $z > 0$, where the inhomogeneity is supposed to be small ($\varepsilon^2 \ll 1$) and the randomness is statistically homogeneous and isotropic. When the wavelength is smaller than the correlation distance a, we may neglect conversion scattering between P and S waves. Then, we can describe the principal characteristics of vector wave propagation by using potentials. For the incidence of plane P-wavelet to the z direction from the homogeneous zone, scalar potential is written as a superposition of harmonic waves of angular frequency ω as $\phi = \int_{-\infty}^{\infty} (2\pi i k_0)^{-1}\, U(\mathbf{x}_\perp, z, \omega)\, \mathrm{e}^{i k_0 z - i\omega t}\, \mathrm{d}\omega$ for $z > 0$, where $\mathbf{x}_\perp = (x, y)$ on the transverse plane. Neglecting the second derivative with respect to z, we have the parabolic-type equation for U as

$$2 i k_0 \partial_z U + \left(\partial_x^2 + \partial_y^2\right) U - 2 k_0^2 \xi(\mathbf{x})\, U = 0\,. \tag{8}$$

We define the two-frequency mutual coherence function (TFMCF) of field U between two different locations on the transverse plane at a distance z and different angular frequencies at ω' and ω'' as $\Gamma_2(\mathbf{x}_{\perp\mathrm{c}}, \mathbf{x}_{\perp\mathrm{d}}, z, \omega_\mathrm{c}, \omega_\mathrm{d}) \equiv \langle U(\mathbf{x}'_\perp, z, \omega') U(\mathbf{x}''_\perp, z, \omega'')^* \rangle$, where ω_c and ω_d are center-of-mass and difference angular frequencies, respectively. In the case of quasi-monochromatic waves $|\omega_\mathrm{d}| \ll |\omega_\mathrm{c}|$, using causality and neglecting back scattering, we derive the master equation for TFMCF. This derivation is called the Markov approximation. For the ith component, the intensity is defined as the ensemble average of the square of displacement $\langle \partial_i \phi\, \partial_i \phi^* \rangle = 1/(2\pi) \int_{-\infty}^{\infty} \widehat{I}_i^{\mathrm{P}}\, \mathrm{d}\omega_\mathrm{c}$. The integrand is the intensity spectral density (ISD) $\widehat{I}_i^{\mathrm{P}}$, which means the time trace of MS amplitude in a band having the central angular frequency ω_c.

Vector Envelopes for a Gaussian ACF

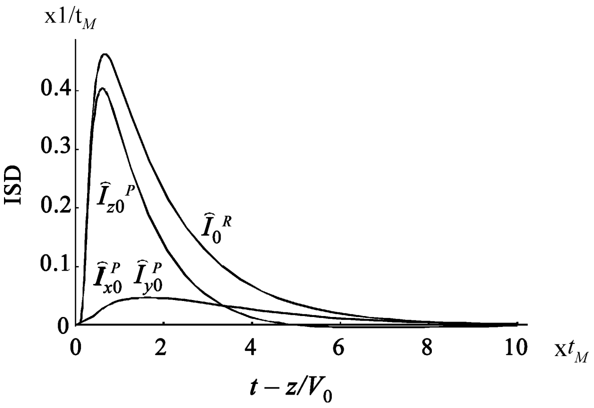
The case of a Gaussian ACF $R(\mathbf{x}) = \mathrm{Exp}\left(-r^2/a^2\right)$ is mathematically tractable. For the initial condition $\widehat{I}_x^{\mathrm{P}} = \widehat{I}_y^{\mathrm{P}} = 0$ and $\widehat{I}_z^{\mathrm{P}} = \delta(t - z/V_0)$ at $z = 0$, ISDs are analytically written as [72]

$$\begin{aligned} \widehat{I}_{x0}^{\mathrm{P}}(z, t, \omega_\mathrm{c}) &= \widehat{I}_{y0}^{\mathrm{P}}(z, t, \omega_\mathrm{c}) \\ &= 2(V_0/z)(t - z/V_0) \cdot \widehat{I}_0^{\mathrm{R}}(z, t, \omega_\mathrm{c}) \\ \widehat{I}_{z0}^{\mathrm{P}}(z, t, \omega_\mathrm{c}) &= [1 - 4(V_0/z)(t - z/V_0)] \\ &\quad \cdot \widehat{I}_0^{\mathrm{R}}(z, t, \omega_\mathrm{c})\ , \end{aligned} \tag{9}$$

where subscript “0” means ISD without the wandering effect. The reference ISD is a solution for scalar waves for the initial condition $\Gamma_2(\mathbf{x}_\perp, z = 0) = 1$ [82]:

$$\begin{aligned} \widehat{I}_0^{\mathrm{R}}(z, t, \omega_\mathrm{c}) &= \frac{1}{t_\mathrm{M}} \frac{\pi}{8} \vartheta_1{}' \left(0, \mathrm{e}^{-\frac{\pi^2}{4} \frac{(t - z/V_0)}{t_\mathrm{M}}}\right) \\ &\quad \cdot H\left(t - \frac{z}{V_0}\right)\ , \end{aligned} \tag{10}$$

where $t_M = \sqrt{\pi}\varepsilon^2 z^2/(2V_0 a)$ is the characteristic time and function ϑ_1' is the derivative of the elliptic theta function of the first kind. Function $\widehat{I}_0^{\mathrm{R}}$ shows a broadened envelope having a delayed peak and a smoothly decaying tail as illustrated by a chained curve in Fig. 7, where solid and broken curves show three-component ISDs $\widehat{I}_{0z}^{\mathrm{P}}$ and $\widehat{I}_{x0}^{\mathrm{P}} (= \widehat{I}_{y0}^{\mathrm{P}})$, respectively, for $V_0 t_\mathrm{M}/z = 0.05$ as an example. All the three component envelopes have broadened traces; however, the peak height of the transverse component is smaller than that of the longitudinal component and the peak delay of the transverse component is larger than that of the longitudinal component. When $\varepsilon^2 z/a \ll 1$, the peak height of $\widehat{I}_{z0}^{\mathrm{P}}$ approximately decays according to the square of travel distance and the peak ratio of the trans-



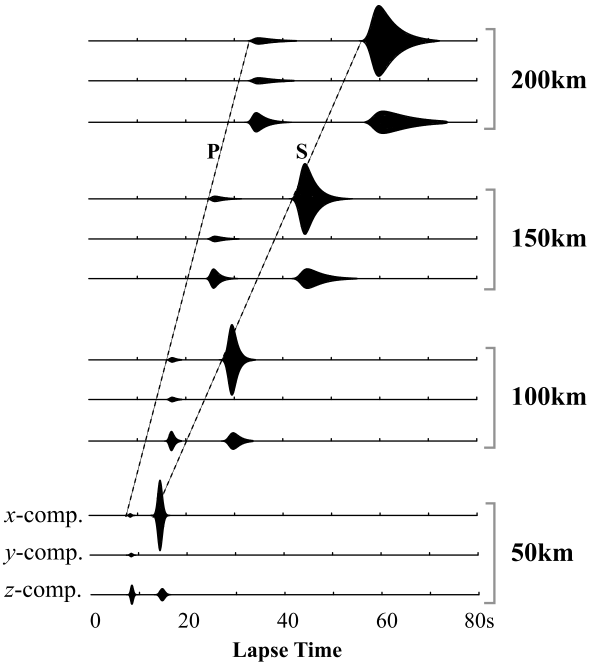
Seismic Waves in Heterogeneous Earth, Scattering of, Figure 7
Chained curve shows the reference ISD without the wandering effect in 3-D random elastic media characterized by a Gaussian ACF for the incidence of a plane P-wavelet. *Solid* and *broken curves* show three-component ISDs without the wandering effect for $V_0 t_M/z = 0.05$. Reproduced from [72]

verse component to longitudinal component is proportional to $\varepsilon^2 z/a$. ISDs $\widehat{I}_x^{\,\mathrm{P}}$, $\widehat{I}_y^{\,\mathrm{P}}$, and $\widehat{I}_z^{\,\mathrm{P}}$ can be calculated by using the convolution of (9) with the travel-time wandering effect $\exp\left[-(V_0 t - z_0)^2/2\sqrt{\pi}\varepsilon^2 az\right] V_0/\sqrt{2\pi\sqrt{\pi}\varepsilon^2 az}$ in time domain. For 2-D cases, the validity of the Markov approximation was numerically confirmed by a comparison with the finite difference simulations [11,36].

The above synthesis can be simply extended to S wave envelopes. Extension from plane wave incidence to impulsive radiation from a point source is also possible [73]. Figure 8 shows simulated three-component RMS envelopes along the z axis for a point source radiation, where random media are characterized by average P and S wave velocities, 6 km/s and 3.46 km/s, respectively, and a Gaussian ACF with $\varepsilon = 5\%$ and $a = 5$ km. We assume that P-wavelet radiation is isotropic and S-wavelet radiation is axially symmetric around the y axis with polarization to the x axis, where the ratio of S to P-wave source energy is chosen to be 23.3. Envelope broadening is common to both P and S waves in the syntheses. Excitation of the transverse component for P-waves and that of the radial component for S-waves are prominent. The appearance of scattered S-waves having long duration in synthesized envelopes at large travel distances qualitatively well explains observed characteristics shown in Fig. 2.

Randomness in the Lithosphere

Applying the envelope broadening model to S-wave seismograms recorded in Kanto, Japan, Sato [70] estimated the ratio $\varepsilon^2/a \approx 10^{-3}\,\mathrm{km}^{-1}$ with the assumption of a Gaussian ACF and S-wave attenuation $Q^{-1} = 0.014 f^{-1}$.



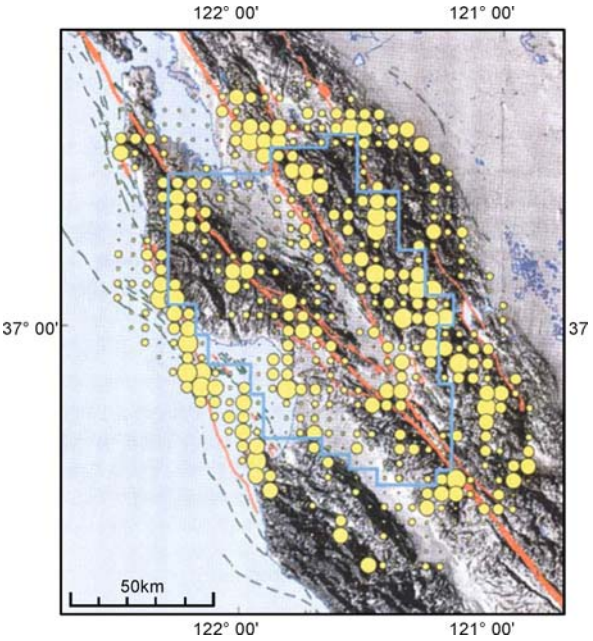
Seismic Waves in Heterogeneous Earth, Scattering of, Figure 8
Synthesized vector envelopes in random media characterized by a Gaussian ACF for radiation of P wavelet and S wavelet with a polarization to the *x*-axis from a point source. Reproduced from [73]

Saito et al. [66] studied the case of a von Kármán-type random media having a power-law spectrum at large wavenumbers, which are more appropriate for the real Earth inhomogeneity. The resultant envelope shows frequency dependence, which is controlled by the roll-off of the PSDF. Analyzing the hypocentral-distance dependence and frequency dependence of S-wave seismogram envelopes in northern Honshu, Japan for 2–32 Hz, Saito et al. [66] estimated parameters of von Kármán-type ACF as $\kappa = 0.6$ and $\varepsilon^{2.2}/a \approx 10^{-3.6}\,\mathrm{km}^{-1}$ with $Q^{-1} = 0.009 f^{-1}$. It means the PSDF decreases as wavenumber to the power of −4.2. Petukhin and Gusev [55] averaged S-wave seismogram envelopes of small earthquakes recorded in Kamchatka and compared the shapes with those numerically calculated for various types of random media. They concluded that random media whose PSDF decreases as the wavenumber to the power of −3.5 to −4 are appropriate.

Spatial Variation of Scattering Characteristics

Scattering Coefficient and Active Faults

Precisely examining coda envelopes of local earthquakes against lapse time measured from the origin time, we find



Seismic Waves in Heterogeneous Earth, Scattering of, Figure 9
Distribution of relative scattering coefficient at a depth of 0–5 km in central California revealed from the coda envelope inversion. *Circles* with larger diameter indicate stronger scattering and *solid lines* represent active faults. Reproduced with permission from [49]

temporal fluctuations around the smoothly decaying master curve. We may interpret that swellings and dips around the master curve are caused by stronger and weaker scatterers, respectively, distributed in the subsurface. By using a single isotropic scattering model, Nishigami [48] proposed an inversion scheme from coda envelopes of local earthquakes recorded at multiple stations for estimating the spatial variation of the scattering coefficient. Applying this inversion scheme to coda records obtained in central California, Nishigami [49] mapped the distribution of relative scattering coefficient in the shallow crust as in Fig. 9. A good correlation is found between sub-parallel active faults and relatively stronger scattering zones marked by larger circles, where some large circles are caused by topographic roughness. He also suggested that segment boundaries of the San Andreas Fault are characterized by relatively stronger scattering.

Stacking forward scattered energy in the coda of teleseismic P waves observed by a local seismographic network, Revenaugh [59] proposed a Kirchhoff coda migration method, which puts a focus on small-angle scattering from the forward direction. He made a map of *P*-wave scatterers in the upper mantle beneath southern California. Between depths of 50 km and 200 km, the southern flank of the slab subducting beneath the Transverse Ranges was marked by strong scattering. Using the same method, Revenaugh [60,61] estimated geographic variation of the statistical significance of scattering potential in the upper crust in California, where the scattering potential is a measure of the likelihood that scattering strength locally exceeds the regional mean. In the region surrounding the 1992 Landers earthquake of M 7.3, he found a noticeable tendency for aftershocks to cluster in regions of strong scattering potential.

There were more precise mappings of scattering coefficient. Slant-stacking records of 12 explosions in the Awaji island, Japan registered by a dense seismic array for a 6–10 Hz band, Matsumoto et al. [45] mapped the spatial distribution of PP single scatterers. The resultant distribution of scatterers shows higher strengths beneath the initiation point of the mainshock rupture and in the southwestern part of the fault plane of the 1995 Kobe earthquake (M 7.2). Analyzing precisely aftershock records of the 2000 western Tottori earthquake (M 7.3), Japan registered by a dense seismic network, Asano and Hasegawa [5] found strong scattering along and around the fault zone of 20 km in length.

Coda Attenuation and Deformation Zone

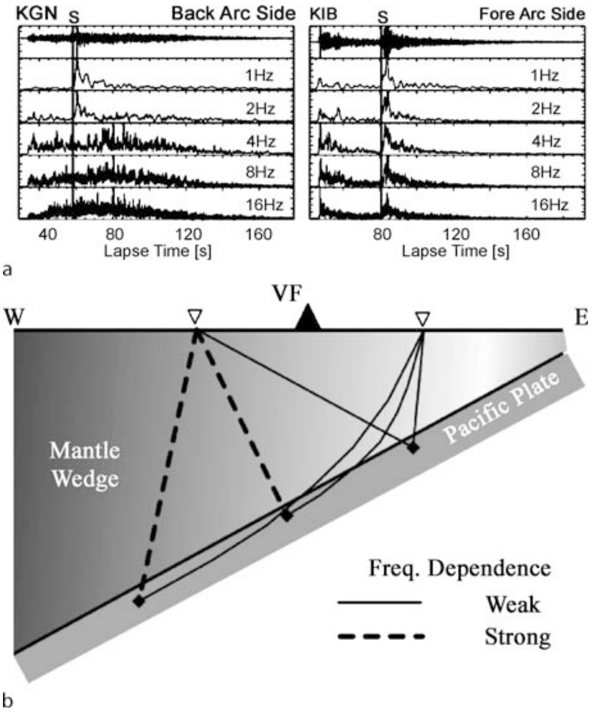
Regional variation of coda attenuation Q_c^{-1} has been measured from the decay gradient of coda amplitude envelopes of small earthquakes in various areas in the world. Jin and Aki [31] made a map of Q_c^{-1} at 1 Hz in China. They reported that Q_c^{-1} is as large as 0.01 in Tibet at the active continental collision. They found that large historical earthquakes took place in large Q_c^{-1} regions. Jin and Aki [33] made precise analysis of Q_c^{-1} for 1–32 Hz in Japan. They found significant spatial variation up to a factor of 3 for the lower frequency bands, as well as its strong frequency dependence. They found conspicuous large Q_c^{-1} for 1–4 Hz in a narrow belt from Niigata towards the south-west to the Biwa lake along the Japan Sea coast, which coincides with a high-deformation rate zone revealed from the GPS observation. For frequency bands 4–16 Hz (2–4 Hz in Kyushu), large Q_c^{-1} areas agree with volcanic and geothermal areas.

Attenuation and Volcanoes

Yoshimoto et al. [91] studied the spatial variation of MS amplitude of *S*-coda at a fixed lapse time across the volcanic front in northeastern Honshu, Japan: *S*-coda energy is uniformly distributed in the fore-arc side, whereas an exponential decrease with horizontal offset to the west from the volcanic front was found in the back-arc side. The decay rate increases with increasing frequency. They interpreted this variation by a diffusion–absorption model, where the intrinsic absorption factor of S-wave $Q_{\mathrm{Int}}^{-1} = 0.02$ at a frequency of 10 Hz beneath the back-arc side, which is about twice as large as those reported for the fore-arc side.

Scattering and Volcanoes

Medium heterogeneity is strong beneath volcanoes. Applying the diffusion model to seismogram envelopes beneath Merapi volcanoes, Friedrich and Wegler [16] estimated the total scattering coefficient as large as 5 km^{-1} as shown in Fig. 4. Nishimura et al. [50] applied an envelope inversion method based on the isotropic scattering model for PP and PS scattering to artificial explosion records obtained in Jemez volcanic field, New Mexico. They found that the mid-crust under most of the region is fairly transparent but that the lower crust is heterogeneous. The strongest scattering occurs at shallow depths beneath the center of the caldera, where the medium is highly heterogeneous.



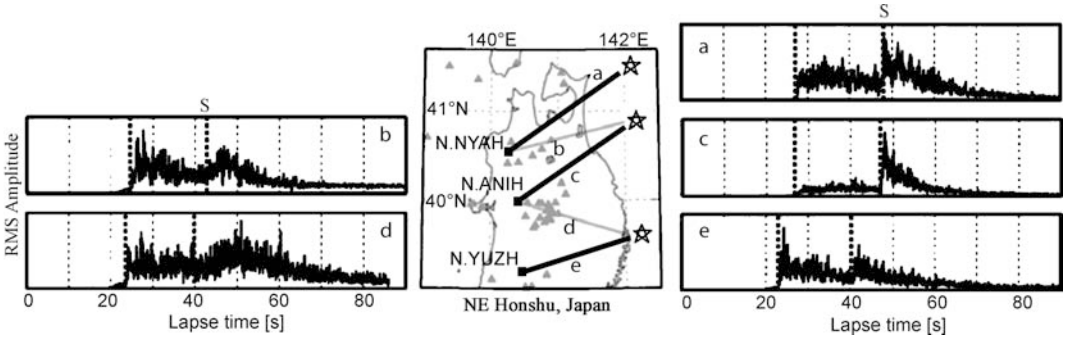
Seismic Waves in Heterogeneous Earth, Scattering of, Figure 10
a Seismogram envelopes observed in the back-arc side and fore-arc side in Kanto-Tokai, Japan and b a schematic illustration of seismic rays. Reproduced from [53]

Obara and Sato [53] analyzed S-wave envelopes of microearthquakes in Kanto-Tokai, Japan, where the Pacific plate is subducting from east to west, to examine regional differences in their envelope broadening. As shown by examples in Fig. 10a, envelope broadening is typically stronger for higher frequencies in records at stations on the back-arc side of the volcanic front but weaker and frequency independent in records on the fore-arc side. These regional differences in the envelope broadening mean that PSDF of velocity inhomogeneity is rich in short-wavelength components in the mantle wedge on the back-arc side and poor on the fore-arc side as schematically illustrated in Fig. 10b. Takahashi et al. [83] precisely examined how the peak delay from the S-wave onset depends on the ray path in northern Japan. They found that peak delays observed in the back-arc side of the volcanic front are larger for rays which propagate beneath Quaternary volcanoes (see Fig. 11b and d); however, peak delays for rays which propagate between Quaternary volcanoes are as short as those observed in the fore-arc side (see Fig. 11a, c, and e). Large peak delay suggests strong scattering due to medium inhomogeneity. That is, the structure beneath Quaternary volcanoes is not only characterized by low velocity and large intrinsic absorption revealed from tomography studies but also by strong inhomogeneity.

Nonisotropic Random Medium Oceanic Slab

If random media are statistically nonisotropic, scattering contribution depends on the propagation direction. Saito [65] studied the envelope broadening in nonisotropic random media based on the Markov approximation. His simulations show that the envelope of scalar wavelet propagating in parallel to the longer correlation direction has longer duration compared to that with the shorter correlation direction. The effective envelope broadening in the elongated direction shows the wave trap phenomenon of nonisotropic random media.

An intensity anomaly is observed on the eastern seaboard of northern Japan for deep earthquakes. The waveform records in the region of high intensity show a low-frequency ($f < 0.25$ Hz) onset for both P and S waves, followed by large-amplitude high-frequency ($f > 2$ Hz) later arrivals with long coda. A simple subduction zone model comprising a high-velocity plate with low attenuation cannot explain quantitatively these observed facts. Furumura and Kennett [17] proposed a scattering slab model that the nonisotropic random structure in the Pacific plate works as a wave-guide for high-frequency seismic waves. Their preferred random medium is characterized by a von Kármán-type ACF with elongated cor-



Seismic Waves in Heterogeneous Earth, Scattering of, Figure 11
Path dependence of RMS envelopes (16–32 Hz) in northeastern Honshu, Japan, where *stars* and *triangles* indicate earthquake epicenters and Quaternary volcanoes, respectively. Ray paths b and d travel beneath Quaternary volcanoes (*triangles*), and a, c, and e travel between Quaternary volcanoes. Reproduced from [83]

relation distance of about 10 km parallel to the plate margin and much shorter correlation length of about 0.5 km in thickness and ε of about 2%. They clearly demonstrated the scattering waveguide effects and frequency selectivity for seismic waves traveling through the Pacific plate by using 3-D numerical simulations.

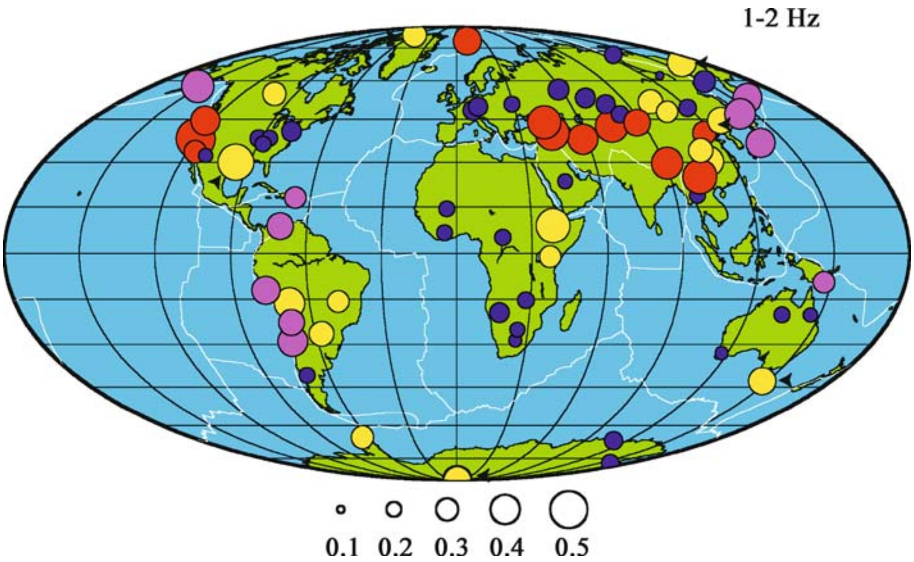
Lateral Variation of Lithospheric Heterogeneity

Korn [34] developed the energy flux model [15] appropriate for the wave front of teleseismic P and P-coda waves propagating through a scattering layer. During the propagation the primary wave loses energy due to scattering and intrinsic absorption, then the scattered energy appears as coda energy behind the wave front. From the analysis of vertical component trace envelopes observed in the world, Korn [35] found strong scattering at island arcs and smaller scattering on stable continental areas like Australia. Nishimura et al. [52] analyzed the energy partition of teleseismic P and P-coda into the transverse component to evaluate the lithospheric heterogeneity in the western Pacific region. They showed the presence of strong heterogeneity in and around the tectonically active regions. Kubanza et al. [37] systematically characterized the medium heterogeneity of the lithosphere by analyzing the partition of P-wave energy into the transverse component for 0.5–4 Hz. They found significant regional differences as shown in Fig. 12. The energy partition to the transverse component is small at stations on stable continents while the partition is large at stations in tectonically active regions such as island arcs or collision zones.

Random Inhomogeneity in the Lithosphere and Mantle

Records of earthquakes registered by arrays of seismographs are useful for the statistical measurement of the Earth inhomogeneity. Aki [2] first analyzed teleseismic P-waves centered on about 0.6 Hz registered by a seismic array in Montana for the quantification of the lithospheric inhomogeneity. Measuring transverse correlation functions of teleseismic P-waves arriving from near vertical incidence, he found a positive correlation between log-amplitude and phase fluctuations as theoretically predicted for a Gaussian ACF. From plots of the ratio of RMS log-amplitude to RMS phase fluctuations against the correlation between log-amplitude and phase fluctuations, he estimated the thickness of the inhomogeneous lithosphere to be 60 km, $a = 10$ km, and $\varepsilon = 4\%$. Flatté and Wu [12] measured the transverse correlation of log-amplitude and phase fluctuations of teleseismic P-wave beams with 2 Hz center frequency recorded at NORSAR. They also introduced the new concept of angular correlation functions, which are based on measurements of two rays with different incident angles. They proposed a model for lithospheric and asthenospheric inhomogeneities that consists of two overlapping layers: the upper layer extending from the surface to about the 200 km depth has a white PSDF, however, the lower layer extending from 15 to 250 km has poor amplitude in the short wavelength spectrum. There are more small-scale inhomogeneities near the surface compared with the deeper portions.

The radiative transfer theory with scattering coefficients calculated from the Born approximation was also used for the study of mantle inhomogeneity. Analyzing stacked P and P coda envelopes of teleseismic (>10°) events at 1 Hz, Shearer and Earle [79] concluded that most scattering occurs in the lithosphere and upper mantle, but that some lower mantle scattering is likely required. They estimated ε to be 3–4% and $a = 4$ km in the upper mantle and 0.5% and 8 km in the lower mantle. Analyzing envelopes of precursors to PKP, Margerin and Nolet [43] found that inhomogeneity can not be restricted to the D″ layer and a small inhomogeneity spread over the whole



Seismic Waves in Heterogeneous Earth, Scattering of, Figure 12
Plots of the square root of the relative partition of wave energy into the transverse component (1–2 Hz) by a diameter of the circle revealed from the teleseismic P-waves. Reproduced from [37]

lower mantle. They proposed a von Kármán ACF of $\kappa = 0$ for random media, which has a power law PSDF rich in short wavelengths compared with an exponential ACF. They mentioned that ε of 0.1–0.2% in the whole lower mantle is enough to explain the observation even though correlation distance is irresolvable because of the limited range of observations.

Imaging of Subsurface Heterogeneity

There have been developments in deterministic imaging of medium inhomogeneity from the analysis of array records of P-coda waves. When the structures of interest are characterized by laterally variable stratification, the receiver-function technique [38] is useful since it is based on the deconvolution of the horizontal component trace in the radial direction by the vertical component trace for measuring the Ps conversion depth. On the other hand, scattering from localized volume inhomogeneity is most readily treated by using the Born approximation. Analyzing array records of teleseismic P coda in central Oregon by using the Born approximation for both forward scattered waves and backscattered free-surface reflected waves, Rondenay et al. [62] successfully imaged the precise structure of the Cascadia subduction zone, which is consistent with the consequences of prograde metamorphic reactions occurring within the oceanic crust. Analyzing array records of P coda waves of regional earthquakes at Izu-Oshima volcano, which erupted in 1986, Mikada et al. [46] deterministically imaged PP and PS scatterers on the basis of diffraction tomography. They interpreted a cloud of scatterers centered at about 10 km depth beneath the volcano crater as a primary magma reservoir and smaller and shallower patches of high scattering strength with sub-magma reservoirs.

Temporal Change in the Earth Medium Structure

There were reports on the temporal changes in the Earth medium structure revealed from the analyses of scattered waves. One is coda amplitude envelope analysis, which gives information about the change in intrinsic absorption and scattering strength of the crustal heterogeneity. Another is coda phase interferometry, which offers information about the change in background velocity.

Change in Coda Characteristics

Monitoring coda envelopes of local earthquakes, Gusev and Lemzikov [23] reported temporal change in Q_c^{-1} before and after the 1971 Ust-Kamchatsk earthquake (M 7.8), and Jin and Aki [30] reported temporal change in Q_c^{-1} associated with the 1976 Tangshan earthquake (M 7.8) in China. Their observation attracted the interest of geophysicists to the temporal variation of coda characteristics because of a potential for monitoring the stress accumulation process preceding an earthquake occurrence. Analyzing high-frequency seismograms recorded at Riverside, California for 55 years, Jin and Aki [32] found a temporal variability in Q_c^{-1} at about 1.6 Hz having a positive correlation with the seismic b-value calculated for $M > 3$ earthquakes within a 180 km radius. The seismic b-value is a measure of the ratio between the numbers of small to large earthquakes; smaller b-values mean that there are relatively fewer small earthquakes compared to the number of larger ones. They interpreted these changes in Q_c^{-1} and the b-value by creep fractures in the ductile part of the lithosphere. Hiramatsu [25] precisely examined the temporal variation in Q_c^{-1} and b-value for 10 years before and after the 1995 Hyogo-ken Nanbu earthquake (M 7.2) in Japan. At frequencies between 1.5 and 4.0 Hz the temporal variation in Q_c^{-1} increased after the mainshock occurrence, where the variation in b-value was opposite.

Sato [69] analyzed the relation between coda duration time and earthquake magnitude of small earthquakes before and after an M 6.8 earthquake in central Japan. He found that coda durations were anomalously longer than usual for 16 months before the earthquake occurrence. From 24-year observation of coda at 0.5 Hz in Kamchatka Gusev [19] reported prominent anomalies in coda level residual from the mean coda excitation level at 100s lapse time associated with two M 8 earthquakes and a volcanic eruption.

Sawazaki et al. [76] measured the temporal variation of the spectral ratio of coda waves registered on the ground surface to that at the bottom of a borehole of 100 m depth in Japan, which experienced strong ground motion of several hundred gals. They reported a sudden drop of the site amplification factor caused by earthquake strong motion and gradual recovery for a few years approaching to the original ratio. They suggested crack formation and ground water movement for explaining the site factor weakening observed.

Coda Interferometry

Pairs of earthquakes with almost identical focal mechanisms are called earthquake doublets. The cross-correlation function of earthquake doublet records allow us to detect differences in the background velocity of the Earth medium that took place in between the pair of earthquakes. Applying the phase spectral analysis to coda wave records of earthquake doublets before and after the 1979 Coyote earthquake of M 5.9 in California, Poupinet et al. [57] found that the coda wave arrivals for some stations are progressively delayed for the second earthquake in the doublet. They interpreted systematic variation along the coda as a decrease of background S-wave velocity by 0.2% in an oblong region 5–10 km in radius at the south end of the aftershock zone. Applying the phase spectral analysis to records of repeated artificial explosions, Nishimura et al. [51] found that the average seismic velocity of the crust in the frequency range of 3–6 Hz decreased by about 1% around the focal region of an M 6.1 earthquake at Iwate volcano in northeastern Honshu, Japan in 1998. They interpreted this velocity drop by the dilatation caused by the M 6.1 earthquake with stress sensitivity of the velocity change $(\delta V/V)/\delta\sigma$ of the order of 0.1 MPa^{-1}. From the set of successive artificial explosion experiments, they observed gradual recovery of the seismic velocity towards its original value over the next four years. While interferometry detected a change in velocity of the order of 1%, it was unidentifiable from travel time analysis of first arrivals. Using coda waves is superior to direct waves since coda waves volumetrically sample the Earth medium.

Snieder et al. [81] demonstrated detection of the nonlinear dependence of the seismic velocity in granite on temperature and the associated acoustic emissions from the interference measurement of coda waves in rock samples as a laboratory experiment. They named this method "coda interferometry" and proposed to use it for detecting the presence of temporal changes in the medium, or in diagnostic mode. There is an idea to retrieve the Green function from the stacked cross-correlation function (CCF) of multiple scattered waves or microseisms at a pair of stations on condition that the propagation directions of those waves are randomly isotropic. Stacking CCFs of coda waves at several pairs of stations for regional earthquakes in Mexico, Campillo and Paul [7] estimated the surface wave velocity between each station pair from the peak delay. The idea was extended for monitoring the temporal change in the crustal structure. Wegler and Sens-Schönfelder [84] computed the ACF of microseisms recorded at a site in the vicinity of the source region of the 2004 Mid-Niigata earthquake (M 6.6) in Japan for three months. They detected a sudden decrease of relative seismic velocity in the crust of 0.6% at the occurrence of the earthquake from the temporal variation of stacked ACFs.

Future Directions

In addition to classic parameterization as a layered structure with sharp edges and smooth velocity perturbation, we introduced new approaches using scattered waves that reflect solid Earth heterogeneity. For high-frequency seismograms of earthquakes, envelope characteristics such as the excitation level and the decay gradient of coda envelopes and the envelope broadening of the direct wavelet are useful for the study of small-scale inhomogeneities. The lithospheric inhomogeneity is phenomenologically well characterized by the scattering coefficient and coda attenuation factor as a function of frequency. Furthermore, the power spectral density function of random velocity inhomogeneity is estimated from the frequency dependence of high-frequency seismogram envelopes of local earthquakes or the array analysis of teleseismic waves. The radiative transfer theory with scattering coefficients calculated from the Born approximation and the Markov approximation for the parabolic wave equation are useful mathematical tools for the analyses.

Scattering characteristics are found to vary spatially reflecting seismotectonic settings. It will be necessary for us to make a classification of seismogram-envelope patterns in various regions in the world under different tectonic conditions. It is interesting to model how such a variation of medium inhomogeneity was created through the geodynamic process. Compared to the lithospheric inhomogeneity, there were insufficient numbers of studies on the mantle inhomogeneity. It will be necessary to map the distribution of inhomogeneities deep in the mantle, which is useful for the study of the evolution of the planet Earth.

For mathematical simplicity; however, most approaches assume homogeneity and isotropy of randomness and a constant background velocity, which are somewhat different from reality. It will be necessary to mathematically develop the envelope synthesis in inhomogeneous media that are a superposition of small-scale random inhomogeneities and a gradually varying background velocity. As revealed from the ray path dependence of S-wave envelope broadening, randomness varies from place to place. It is also necessary to develop the envelope synthesis for random media having spatially varying statistical parameters. In addition, it is important to examine how conversion scattering between P and S waves contributes to form spindle-like envelopes in highly scattering media as shown in high-frequency seismograms observed in volcanoes and on the Moon.

For further understanding, there are monographs that treat the discussed subjects as follows: Sato and Fehler [74] review seismological observation and mathematical models; Shapiro and Hubral [78] put special focus on wave propagation through stratified random media; Goff and Holliger [18] summarize the crustal heterogeneity; Wu and Maupin [86] compile recent developments in mathematical modeling of wave propagation in inhomogeneous media; Chandrasekhar [8] is a classic text for radiative transfer theory; Ishimaru [29] and Rytov et al. [63] offer advanced mathematical tools for the study of wave propagation in random media.

Acknowledgments

The author thanks Michael Korn, Heiner Igel, and an anonymous reviewer. Their comments were helpful for improving the readability. The author is grateful to NIED, Japan for providing digital seismic data.

Bibliography

1. Aki K (1969) Analysis of seismic coda of local earthquakes as scattered waves. J Geophys Res 74:615–631
2. Aki K (1973) Scattering of P waves under the Montana LASA. J Geophys Res 78:1334–1346
3. Aki K (1980) Attenuation of shear-waves in the lithosphere for frequencies from 0.05 to 25 Hz. Phys Earth Planet Inter 21:50–60
4. Aki K, Chouet B (1975) Origin of coda waves: Source, attenuation and scattering effects. J Geophys Res 80:3322–3342
5. Asano Y, Hasegawa A (2004) Imaging the fault zones of the 2000 western Tottori earthquake by a new inversion method to estimate three-dimensional distribution of the scattering coefficient. J Geophys Res 109:B06306. doi:10.1029/2003JB002761
6. Atkinson GM (1993) Notes on ground motion parameters for Eastern North America: Duration and H/V ratio. Bull Seismol Soc Am 83:587–596
7. Campillo M, Paul A (2003) Long-Range Correlations in the Diffuse Seismic Coda. Science 299:547–549. doi:10.1126/science.1078551
8. Chandrasekhar S (1960) Radiative Transfer. Dover, New York
9. Dainty AM, Toksöz MN (1981) Seismic codas on the earth and the moon: A comparison. Phys Earth Planet Inter 26:250–260
10. Fehler M, Hoshiba M, Sato H, Obara K (1992) Separation of scattering and intrinsic attenuation for the Kanto-Tokai region, Japan, using measurements of S-wave energy versus hypocentral distance. Geophys J Int 108:787–800
11. Fehler M, Sato H, Huang LJ (2000) Envelope broadening of outgoing waves in 2-D random media: A comparison between the Markov approximation and numerical simulations. Bull Seismol Soc Amer 90:914–928
12. Flatté SM, Wu RS (1988) Small-scale structure in the lithosphere and asthenosphere deduced from arrival time and amplitude fluctuations at NORSAR. J Geophys Res 93:6601–6614
13. Foldy LL (1945) The multiple scattering of waves- I General theory of isotropic scattering by randomly distributed scatterers. Phys Rev 67:107–119
14. Frankel A, Clayton RW (1986) Finite difference simulations of seismic scattering: Implications for the propagation of short-period seismic waves in the crust and models of crustal heterogeneity. J Geophys Res 91:6465–6489
15. Frankel A, Wennerberg L (1987) Energy-flux model of seismic coda: Separation of scattering and intrinsic attenuation. Bull Seismol Soc Am 77:1223–1251
16. Friedrich C, Wegler U (2005) Localization of seismic coda at Merapi volcano (Indonesia). Geophys Res Lett 32:L14312. doi:10.1029/2005GL023111
17. Furumura T, Kennett BLN (2005) Subduction zone guided waves and the heterogeneity structure of the subducted plate: intensity anomalies in northern Japan. J Geophys Res 110:B10302. doi:10.1029/2004JB003486

18. Goff JA, Holliger K (2002) Heterogeneity in the Crust and Upper Mantle – Nature, Scaling and Seismic Properties. Kluwer Academic/Plenum Publishers, Dorderecht, pp 1–358
19. Gusev AA (1995) Baylike and continuous variations of the relative level of the late coda during 24 years of observation on Kamchatka. J Geophys Res 100:20311–20319
20. Gusev AA (1995) Vertical profile of turbidity and coda Q. Geophys J Int 123:665–672
21. Gusev AA, Abubakirov IR (1987) Monte-Carlo simulation of record envelope of a near earthquake. Phys Earth Planet Inter 49:30–36
22. Gusev AA, Abubakirov IR (1996) Simulated envelopes of non-isotropically scattered body waves as compared to observed ones: Another manifestation of fractal heterogeneity. Geophys J Int 127:49–60
23. Gusev AA, Lemzikov VK (1985) Properties of scattered elastic waves in the lithosphere of Kamchatka: Parameters and temporal variations. Tectonophysics 112:137–153
24. Hemmer PC (1961) On a generalization of Smoluchowski's diffusion equation. Physica A 27:79–82
25. Hiramatsu Y, Hayashi N, Furumoto M (2000) Temporal changes in coda *Q*21 and *b* value due to the static stress change associated with the 1995 Hyogo-ken Nanbu earthquake. J Geophys Res 105:6141–6151
26. Holliger K, Levander A (1992) A stochastic view of lower crustal fabric based on evidence from the Ivrea zone. Geophys Res Lett 19:1153–1156
27. Hoshiba M (1991) Simulation of multiple-scattered coda wave excitation based on the energy conservation law. Phys Earth Planet Inter 67:123–136
28. Hoshiba M, Sato H, Fehler M (1991) Numerical basis of the separation of scattering and intrinsic absorption from full seismogram envelope – A Monte-Carlo simulation of multiple isotropic scattering. Pa Meteorol Geophys, Meteorol Res Inst 42:65–91
29. Ishimaru A (1978) Wave Propagation and Scattering in Random Media, vol 1 and 2. Academic, San Diego
30. Jin A, Aki K (1986) Temporal change in coda *Q* before the Tangshan earthquake of 1976 and the Haicheng earthquake of 1975. J Geophys Res 91:665–673
31. Jin A, Aki K (1988) Spatial and temporal correlation between coda *Q* and seismicity in China. Bull Seismol Soc Am 78:741–769
32. Jin A, Aki K (1989) Spatial and temporal correlation between coda Q^{-1} and seismicity and its physical mechanism. J Geophys Res 94:14041–14059
33. Jin A, Aki K (2005) High-resolution maps of Coda Q in Japan and their interpretation by the brittle-ductile interaction hypothesis. Earth Planets Space 57:403–409
34. Korn M (1990) A modified energy flux model for lithospheric scattering of teleseismic body waves. Geophys J Int 102:165–175
35. Korn M (1993) Determination of site-dependent scattering Q from P-wave coda analysis with an energy-flux model. Geophys J Int 113:54–72
36. Korn M, Sato H (2005) Synthesis of plane vector-wave envelopes in 2-D random elastic media based on the Markov approximation and comparison with finite difference simulations. Geophys J Int 161:839–848
37. Kubanza M, Nishimura T, Sato H (2006) Spatial variation of lithospheric heterogeneity on the globe as revealed from transverse amplitudes of short-period teleseismic P-waves. Earth Planets Space 58:45–e48
38. Langston CA (1979) Structure under Mount Rainer, Washington, inferred from teleseismic body waves. J Geophys Res 84:4749–4762
39. Larose E, Margerin L, van Tiggelen BA, Campillo M (2004) Weak Localization of Seismic Waves. Phys Rev Lett 93:048501-4. doi:10.1103/PhysRevLett.93.048501
40. Lee WS, Sato H, Lee KW (2003) Estimation of S-wave scattering coefficient in the mantle from envelope characteristics before and after the ScS arrival. Geophys Res Lett 30:2248. doi:10.1029/2003GL018413
41. Margerin L (2005) Introduction to radiative transfer of seismic waves. In: Levander A, Nolet G (eds) Seismic Earth: Array Analysis of Broad-band Seismograms, Geophysical Monograph Series, vol 157, chap 14. AGU, Washington, pp 229–252
42. Margerin L, Campillo M, van Tiggelen BA (2001) Coherent backscattering of acoustic waves in the near field. Geophys J Int 145:593–603
43. Margerin L, Nolet G (2003) Multiple scattering of high-frequency seismic waves in the deep Earth: PKP precursor analysis and inversion for mantle granularity. J Geophys Res 108, B11:2514. doi:10.1029/2003JB002455
44. Matsumoto S (2005) Scatterer density estimation in the crust by seismic array processing. Geophys J Int 163:622–628
45. Matsumoto S, Obara K, Hasegawa A(1998) Imaging P-wave scatterer distribution in the focal area of the 1995 M7.2 Hyogo-ken Nanbu (Kobe) Earthquake. Geophys Res Lett 25:1439–1442
46. Mikada H,Watanabe H, Sakashita S (1997) Evidence for subsurface magma bodies beneath Izu-Oshima volcano inferred from a seismic scattering analysis and possible interpretation of the magma plumbing system of the 1986 eruptive activity. Phys Earth Planet Inter 104:257–269
47. Nakahara H, Nishimura T, Sato H, Ohtake M (1998) Seismogram envelope inversion for the spatial distribution of high-frequency energy radiation from the earthquake fault: Application to the 1994 far east off Sanriku earthquake, Japan. J Geophys Res 103:855–867
48. Nishigami K (1991) A new inversion method of coda waveforms to determine spatial distribution of coda scatterers in the crust and uppermost mantle. Geophys Res Lett 18:2225–2228
49. Nishigami K (2000) Deep crustal heterogeneity along and around the San Andreas fault system in central California and its relation to the segmentation. J Geophys Res 105:7983–7998
50. Nishimura T, Fehler M, Baldridge WS, Roberts P, Steck L (1997) Heterogeneous structure around the Jemez Volcanic Field, New Mexico, USA, as inferred from the envelope inversion of active-experiment seismic data. Geophys J Int 131:667–681
51. Nishimura T, Tanaka S, Yamawaki T, Yamamoto H, Sano T, Sato M, Nakahara H, Uchida N, Hori S, Sato H (2005) Temporal changes in seismic velocity of the crust around Iwate volcano, Japan, as inferred from analyses of repeated active seismic experiment data from 1998 to 2003. Earth Planets Space 57:491–505
52. Nishimura T, Yoshimoto K, Ohtaki T, Kanjo K, Purwana I (2002) Spatial distribution of lateral heterogeneity in the upper mantle around the western Pacific region as inferred from analysis of transverse components of teleseismic P-coda. Geophys Res Lett 29:2089–2137. doi:10.1029/2002GL015606

53. Obara K, Sato H (1995) Regional differences of random inhomogeneities around the volcanic front in the Kanto-Tokai area, Japan, revealed form the broadening of S wave seismogram envelopes. J Geophys Res 100:2103–2121
54. Paaschens JCJ (1997) Solution of the time-dependent Boltzmann equation. Phys Rev E 56:1135–1141
55. Petukhin AG, Gusev AA (2003) The Duration-distance Relationship and Average Envelope Shapes of Small Kamchatka Earthquakes. Pure Appl Geophys 160:1717–1743
56. Phillips WS, Aki K (1986) Amplification of coda waves from local earthquakes in Central California. Bull Seismol Soc 76:627–648
57. Poupinet G, Ellsworth WL, Frechet J (1984) Monitoring velocity variations in the crust using earthquake doublets: An application to the Calaveras fault, California. J Geophys Res 89:5719–5731
58. Przybilla J, Korn M, Wegler U (2006) Radiative transfer of elastic waves versus finite difference simulations in two-dimensional random media. J Geophys Res 111:B04305. doi:10.1029/2005JB003952
59. Revenaugh J (1995) A scattered-wave image of subduction beneath the Transverse Ranges. Science 268:1888–1892
60. Revenaugh J (1995) Relationship of the 1992 Landers, California, earthquake sequence to seismic scattering. Science 270:1344–1347
61. Revenaugh J (1999) Geologic Applications of Seismic Scattering. Annu Rev Earth Planet Sci 27:55–73
62. Rondenay S, Bostock MG, Shragge J (2001) Multiparameter two-dimensional inversion of scattered teleseismic body waves 3. Application to the Cascadia 1993 data set. J Geophys Res 106:30795–30807
63. Rytov SM, Kravtsov YA, Tatarskii VI (1987) Principles of Statistical Radio Physics, vol 4, Wave Propagation Through Random Media. Springer, Berlin
64. Ryzhik LV, Papanicolaou GC, Keller JB (1996) Transport equations for elastic and other waves in random media. Wave Motion 24:327–370
65. Saito T (2006) Synthesis of scalar-wave envelopes in two-dimensional weakly anisotropic random media by using the Markov approximation. Geophys J Int 165:501–515. doi:10.1111/j.1365-246X2006.02896.x
66. Saito T, Sato H, Ohtake M (2002) Envelope broadening of spherically outgoing waves in three-dimensional random media having power-law spectra. J Geophys Res 107:2089. doi:10.1029/2001JB000264
67. Sato H (1977) Single isotropic scattering model including wave conversions: Simple theoretical model of the short period body wave propagation. J Phys Earth 25:163–176
68. Sato H (1984) Attenuation and envelope formation of three-component seismograms of small local earthquakes in randomly inhomogeneous lithosphere. J Geophys Res 89:1221–1241
69. Sato H (1987) A precursor-like change in coda excitation before the western Nagano earthquake (Ms = 6.8) of 1984 in central Japan. J Geophys Res 92:1356–1360
70. Sato H (1989) Broadening of seismogram envelopes in the randomly inhomogeneous lithosphere based on the parabolic approximation: Southeastern Honshu, Japan. J Geophys Res 94:17735–17747
71. Sato H (1990) Unified approach to amplitude attenuation and coda excitation in the randomly inhomogeneous lithosphere. Pure Appl Geophys 132:93–121
72. Sato H (2006) Synthesis of vector wave envelopes in three-dimensional random elastic media characterized by a Gaussian autocorrelation function based on the Markov approximation: Plane wave case. J Geophys Res 111:B06306. doi:10.1029/2005JB004036
73. Sato H (2007) Synthesis of vector-wave envelopes in 3-D random elastic media characterized by a Gaussian autocorrelation function based on the Markov approximation: Spherical wave case. J Geophys Res Solid Earth 112:B01301. doi:10.1029/2006JB004437
74. Sato H, Fehler M (1998) Seismic Wave Propagation and Scattering in the Heterogeneous Earth. AIP Press/Springer, New York
75. Sato H, Nakahara H, Ohtake M (1997) Synthesis of scattered energy density for non-spherical radiation from a point shear dislocation source based on the radiative transfer theory. Phys Earth Planet Inter 104:1–281
76. Sawazaki K, Sato H, Nakahara H, Nishimura T (2006) Temporal Change in Site Response Caused by Earthquake Strong Motion as Revealed from Coda Spectral Ratio Measurement. Geophys Res Lett 33:L21303. doi:10.1029/2006GL027938
77. Shang T, Gao L (1988) Transportation theory of multiple scattering and its application to seismic coda waves of impulsive source. Sci Sin 31B:1503–1514
78. Shapiro SA, Hubral P (1999) Elastic Waves in Random Media – Fundamentals of Seismic Stratigraphic Filtering. Springer, Berlin
79. Shearer PM, Earle PS (2004) The global short-period wavefield modeled with a Monte Carlo seismic phonon method. Geophys J Int 158:1103–1117
80. Shiomi K, Sato H, Ohtake M (1997) Broad-band power-law spectra of well-log data in Japan. Geophys J Int 130:57–64
81. Snieder R, Gret A, Douma A, Scales J (2002) Coda wave interferometry for estimating nonlinear behavior in seismic elocity. Science 295:2253–2255
82. Sreenivasiah I, Ishimaru A, Hong ST (1976) Two-frequency mutual coherence function and pulse propagation in a random medium: An analytic solution to the plane wave case. Radio Sci 11:775–778
83. Takahashi T, Sato H, Nishimura T, Obara K (2006) Strong inhomogeneity beneath Quaternary volcanoes revealed from the peak delay analysis of S-wave seismograms of microearthquakes in northeastern, Japan. Geophys J Int 168:90–99. doi:10.1111/j.1365-246X2006.03197.x
84. Wegler U, Sens-Schönfelder C (2007) Fault zone monitoring with passive image interferometry. Geophys J Int 168:1029-1033. doi:10.1111/j.1365-246X2006.03284.x
85. Wu RS (1985) Multiple scattering and energy transfer of seismic waves – separation of scattering effect from intrinsic attenuation – I Theoretical modeling. Geophys J R Astron Soc 82:57–80
86. Wu RS, Maupin V (eds) (2007) Advances in Wave Propagation in Heterogeneous Earth. In: Dmowska R (ed) Advanced in Geophysics, vol 48. Academic Press, San Diego, pp 561–596
87. Yomogida K, Benites R (1995) Relation between direct wave *Q* and coda *Q*: A numerical approach. Geophys J Int 123:471–483
88. Yoshimoto K (2000) Monte-Carlo simulation of seismogram envelope in scattering media. J Geophys Res 105:6153–6161
89. Yoshimoto K, Sato H, Ohtake M (1993) Frequency-dependent attenuation of P and S waves in the Kanto area, Japan, based on the coda-normalization method. Geophys J Int 114:165–174

90. Yoshimoto K, Sato H, Ohtake M (1997) Short-wavelength crustal inhomogeneities in the Nikko area, central Japan, revealed from the three-component seismogram envelope analysis. Phys Earth Planet Inter 104:63–73
91. Yoshimoto K, Wegler U, Korn M (2006) A volcanic front as a boundary of seismic attenuation structures in northeastern Honshu, Japan. Bull Seismol Soc Am 96:637–646
92. Zeng Y, Su F, Aki K (1991) Scattering wave energy propagation in a random isotropic scattering medium I Theory. J Geophys Res 96:607–619
93. Zhao D, Hasegawa A, Horiuchi S (1992) Tomographic imaging of P and S wave velocity structure beneath Northeastern Japan. J Geophys Res 97:19909–19928

Self-assembled Materials

Aatto Laaksonen[1], Lennart Bergström[2]

[1] Division of Physical Chemistry, Dept. of Physical, Inorganic and Structural Chemistry, Arrhenius Laboratory, Stockholm University, Stockholm, Sweden

[2] Materials Chemistry Research Group, Dept. of Physical, Inorganic and Structural Chemistry, Arrhenius Laboratory, Stockholm University, Stockholm, Sweden

Article Outline

Glossary

AI Ab initio
AM1 Austin model 1
AO Atomic orbital
B3LYP Becke's three parameter hybrid LYP functional
BKL Bortz–Kalos–Lebowitz
BO Born–Oppenheimer
CA Cellular automata
CE Cluster expansion
CC Coupled cluster
CI Configuration interaction
CP Car-Parrinello
DCA Dynamic cellular automata
DFT Density functional theory
DLVO Derjaguin–Landau–Verwey–Overbeek
GGA Generalized gradient approximation
HF 1) Hartree–Fock 2) Harris–Foulkes functional
kMC kinetic Monte Carlo
KS Kohn–Sham
LbL Layer-by-layer
LCAO Linear Combination of Atomic Orbitals
LDA Local density approximation
LYP Lee–Young–Parr
MC Monte Carlo
MD Molecular Dynamics
MNDO Modified neglect of differential overlap
MO Molecular Orbital
MP Møller–Plesset
NDDO Neglect of diatomic differential overlap
NDO Neglect of differential overlap
NPT Isobaric ensemble: constant N, P and T
NVE Microcanonical ensemble: constant N, V and E
NVT Canonical ensemble: constant N, V and T
PBC Periodic boundary conditions
PF Phase-field
PM3 Parametric model 3
PP Pseudo-potential
PW Perdew–Wang
QCA Quantum-dot cellular automata
RSA Random sequential adsorption
SA Self-assembly
SAM Self-assembled monolayers
SCF Self-consistent field
STM Scanning tunneling microscopy
TB Tight-binding
XC Exchange correlation

Definition of the Subject

The concept of self-assembly is today highly popular and frequently used to describe a wide range of phenomena. It is also a concept that has a possibility to change the way we produce various types of materials [1]. Self-assembly can broadly speaking be defined as a process with the following features [2], where: (i) it involves pre-existing components, i. e. the components are not formed by the reaction itself. (ii) the process should be reversible to some extent. (iii) it can be controlled by design.

The supramolecular chemistry approach, pioneered by Jean-Marie Lehn [3], where molecular recognition is used to assemble supramolecular materials is clearly an important starting point for much of this work. The pre-existing components are usually molecules, amphiphilic species, or oligomeric or polymeric species. However, self-assembly can also be observed on larger length scales involving e. g. nanoparticles (superlattices) and larger colloidal particles (colloidal crystals). The final self-assembled material should display certain features that can be related to the building blocks in addition to the novel features that are created by the self-assembled material itself. Reversibility is a key feature since self-assembly processes commonly scan a rich energy landscape with many metastable states. This is commonly regulated by the balance between the attractive and repulsive interactions between the self-assembling components, involving e. g. van der Waals, electrostatic, hydrophobic, hydrogen bonding and entropic forces (molecules and mesoscale objects), and magnetic, capillary, and gravity forces (meso- or macroscale objects).

The degree of versatility, simplicity and flexibility of the various self-assembly methods and the ability to introduce specific functions with a high degree of spatial accuracy are important features for self-assembled materials to move beyond academic beauty and be interesting for applications. The preparation of self-assembled coatings can be traced back to the seminal work by Irving Langmuir and Katherine Blodgett prior to the Second World War. The development of a theory for adsorption on surfaces by Langmuir together with the design of a technique that both could measure the pressure-area isotherms of monolayers of amphiphilic molecules assembled at the air-liquid interface, and also be used to transfer these monolayers onto different substrates was a very important discovery that has had a profound effect on future work on self-assembled layers. The observation by Nuzzo and coworkers [4] that alkanethiolates can self-assemble on gold surfaces and result in well-ordered films sparked a large research interest in these so called self-assembled monolayers (SAMs). The relatively non-local nature of the thiol bond to the gold surface allows the much weaker, attractive interaction *between* the amphiphilic chains in the monolayer to control the film structure. The SAMs are a beautiful example of self-assembly as a crucial process in the formation of a dense monolayer.

Self-organization of nanoparticles into two- and three-dimensional superlattices has attracted much interest since the early work on iron oxide "super crystals" [5]. Indeed, understanding and optimizing the structures, at all length scales, of nanocrystal superlattices is an important step towards controlled design of novel nanostructured materials and devices. This work did not really gain momentum until methods to achieve shape and size control with a high fidelity was established in the end of 1980s by e. g. Brus, Steigerwald et al. [6] and Moerup, Thölen and Koch [5] mentioned above. The ability to assemble different nanocrystals with size-tunable optical, electronic and magnetic properties into well-defined structures opens up the possibility to study and develop new materials with tailored couplings between the constituent units [7,8]. Layer-by-layer (LbL) self-assembly of charged polymers has evolved as probably the most versatile technique to create multifunctional coatings on a wide range of substrates [9]. These are but a limited set of examples of self-assembly processes to produce novel materials that has been demonstrated in the last decades [10,11,12,13, 14,15,16,17,18,19,20]. While the examples of self-assembly are plentiful, the attempts to model the self-assembly processes and systems is still in its infancy. This review attempts to be a guide and introduction to the field of computer modeling of self-assembly.

Introduction

Computer modelling of structural properties and dynamical processes in matter is based on physical and mathematical models. Ideally, we would prefer to use first-principles methods of quantum mechanics with no other input than atomic nuclei and the electrons spontaneously forming atoms, molecules and more complicated systems through local intermolecular interactions. Unfortunately this is possible only for systems of very modest sizes using rather drastic approximations and simplifications. We can always blame the computers for not been powerful enough but the truth is that the "first-principles" models would grow too complicated and out of hands very rapidly. We therefore use different physical models to describe matter at different length and time scales together with an increasing amount of empirical data as input to the models.

Atomistic computer simulations such as the classical Molecular Dynamics (MD) are currently run for molecular systems consisting of the order of 10^5 atoms, corresponding typically to a system size (length scale) of 5–10 nm. These simulations are typically extended to 10–100 ns (time scale). This may be enough for simulations of isotropic liquids of simple molecules, whereas for many complex systems, the requirements for the length and time scales are much more extensive. To extend the simulations to cover longer scales requires either simplifications of used models or choosing more suitable modelling technique [21]. In the first strategy the resolution of the model is reduced by removing certain degrees of freedom of the studied system which are not important or can be considered as fluctuations while keeping those degrees of freedom which are operational on the longer scales. This is

Self-assembled Materials, Table 1
Time and length scales covered by modelling methods reviewed in this article

Scale	Typical length	Typical time
sub-atomic	< 0.1 nm	~ 1 fs
atomic	0.1–1 nm	1 fs–1 ns
meso	1 nm–1 μm	1 ns–1 ms
macro	1 μm–	1 ms–

generally called coarse-graining of the model leading to a much fewer interaction sites thereby speeding up the calculations. An alternative approach is to use a hybrid by keeping an accurate description for a selected area of importance and the environment is merely made to a fluctuating medium with more or less specific interactions. These techniques may include combining quantum and classical mechanics [22] or classical mechanics and hydrodynamic/continuum models.

In Table 1 there is a somewhat artificial division of matter on scales from sub-atomic (with nuclei and electrons) to macroscopic. There are modelling and simulation methods designed separately to these scales. Going from one scale to the next in the table requires either coarse-graining or fine-graining of the used model by building a bridge between two techniques. This means taking as much vital information as possible to the next level. Coarse-graining is obviously easier than fine-graining as the latter requires re-activating non-existing degrees of freedom from an under-defined system. Multiscale modelling is currently a hot topic and in a rapid development with the purpose to bridge several scales in a more systematic way.

The aim of this chapter is to describe a box of tools containing some selected modelling methods to study self-assembly. The tools should together cover the scales in the table in Table 1. We will give a short description of each of them together with some examples of applications. At the end of the chapter we show examples of how to combine several techniques to extend length and time scales in modelling. We will work from the "bottom-up" by starting with first-principles calculations on electrons and nuclei. Classical Molecular Dynamics (MD) and Monte Carlo (MC) are mainly used in atomistic simulations with empirically parameterized potential functions. Meso-scale simulations of aggregates and soft particles use coarse-grained models. We will discuss methods equally applicable on both short and long scales such as the kinetic Monte Carlo (kMC) and Phase Field (PF) modelling. All these techniques so far are examples of off-lattice modelling methods. We then discuss two lattice methods with some further development towards more off-lattice character, namely Random Sequential Adsorption (RSA) and Cellular Automata (CA). Finally we will discuss multiscale modeling before the paper ends with a discussion on future directions.

First-Principles Calculations and Simulations

Down in the bottom of the world of atoms and molecules, quantum mechanics should be used to treat the moving nuclei and electrons building up the matter. The electronic structure in atoms and molecules and molecules gives the first level of properties to all matter around us. Theoretical calculations utilize the so-called Born–Oppenheimer (BO) approximations which keeps the much heavier nuclei in fixed positions while the electrons adapt themselves pair wise in binding and non-binding molecular orbitals (MO), constructed as linear combinations of atomic orbitals (LCAO) using suitable basis sets (normally Gaussian functions or plane waves). Computational quantum chemistry is largely based on solving the Schrödinger equations interactively within the self-consistence field (SCF) theory until suitable convergence criteria are reached. The previously commonly used Hartree–Fock (HF) method was made computationally convenient in matrix form by Roothaan [23] in early 50s and allowed accurate calculations for larger and larger molecules at the same time as computers were made more and more powerful [24]. As there is no exact solution to many-body problems ($N > 2$) all modelling methods use approximations to reduce the problem to independent particle or pair problem [25]. Within the HF approximation the full N-electron problem is reduced to an independent N single-electrons moving in a mean-field of the other electrons. By increasing the size of the basis set a so called HF limit can be reached where the electron-electron repulsions are treated uncorrelated. As the electron correlation effects are very important for many molecular properties a number of post-HF schemes have been developed, for example Møller–Plesset (MPn ($n = 2, 4, \ldots$)), configuration interaction (CI) or coupled cluster (CC) [25].

The most affordable quantum chemistry methods are the semi-empirical MO methods [26,27,28]. Among the most popular are those based on the neglect of diatomic differential overlap (NDDO) approximation, such as Austin Model 1 (AM1) [29], Parametric Model 3 (PM3) [30,31] and modified NDO (MNDO) [32]. All these schemes start from HF-Roothaan equations [25] introducing the frozen-core approximation and using only the valence AOs. Some electron-electron integrals are simply neglected, some calculated approximately and some are replaced by empirical parameters. These methods are very fast compared to *ab initio* methods but give often unreliable results; the errors tend to be unsystematic and therefore both relative energies and conformational energies become less reliable when calculated using semi-empirical schemes. Also hydrogen bonds are not accurately described. These shortcomings are found to be largely due to an unbalanced treatment of core-electron and electron-electron interactions. Orthogonalization corrections have been worked out to improve the NDDO schemes [33,34] but have given only a slight improvement. This unbalance is corrected in the recently published Extended NDDO scheme [35] where the old NDDO is expressed as zeroth order term. However, it still remains to be parameterized and packaged to a new semi-empirical scheme, which is a major undertaking.

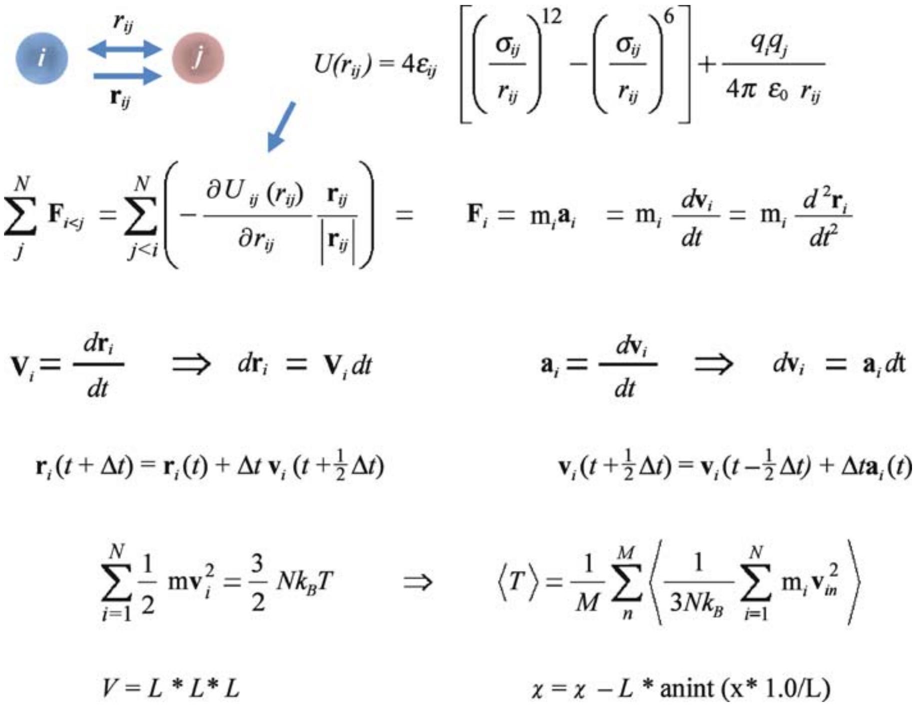
In both materials science and chemistry, the density functional theory (DFT) has recently gained the position as the most popular method [25]. It is basically as fast as the HF method but can incorporate a large part of the electron correlation effects. In DFT the energy is given as a functional of the electron density $\rho(\mathbf{r})$. The contributions to total energy come from the kinetic energy of the electrons, nuclear-electron attraction, electron-electron Coulombic repulsion and finally from combined exchange and correlation energy. The most difficult problem in DFT is to derive the exchange-correlation (XC) term. In materials science for solids and metals the so called local density approximation (LDA) is used as it is suitable in cases of slowly fluctuating electron densities [25]. LDA is not a good choice for chemical problems and so called gradient corrected methods like the generalized gradient approximation (GGA) are used where the XC term is not just functional of the electron density but also derivatives of the density with respect of the coordinates (x, y, z) [25]. The most popular XC functionals are currently those by Lee, Yang and Parr (LYP) [36] and by Perdew and Wang (PW91) [37]. In chemical applications the hybrid methods where the HF exchange energy is mixed the corresponding DFT term have gained popularity. Becke's three parameter hybrid [38] with the LYP functional, the so called B3LYP, is frequently used. Current DFT methods are based on formulation proposed by Kohn and Sham (KS) [39,40] by constructing the total electron density from a KS orbital. This gives conceptually similar equations as in HF method. Car and Parrinello (CP) presented a DFT based Molecular Dynamics method using plane-wave basis functions [41] which is the most popular *ab initio* simulation tool both in chemistry and materials science.

Recently a robust approximate DFT method is introduced where the electron-electron interactions in the total energy density functional in DFT are expanded with respect to a reference density and keeping the first order correction and neglecting the higher order corrections. This is called Harris–Foulkes functional [42,43] and is stable with respect to the used reference density as a small change in the reference density introduces an error only to second order term. Computational schemes based on Harris–Foulkes functionals are normally called *ab initio* tight-binding density functional theory (AB-TBDFT) methods, and were first introduced by Sankey et al. [44]. A highly efficient AB-TBDFT method is recently presented by Tu and Laaksonen [45,46] applicable on simulations of large systems. Various methods to incorporate quantum mechanics into molecular dynamics simulations are reviewed in [22].

First principles electronic structure calculations and simulations have a firm position in materials science. First-principles MD simulations to materials properties are reviewed in [47]. In the following a few examples are chosen from literature. Materials simulations using VASP gives a quantum perspective to materials science according to [48] using pseudo potentials (PP) and a plane wave basis set based on a finite-temperature local-density approximation. Density functional theory meeting statistical physics [49]: from the atomistic to the mesoscopic properties of alloys by combining DFT calculations with the so-called Cluster Expansion (CE) methods and Monte-Carlo (MC) simulations is reported. Determination of solid-state nanostructure from *ab initio* structure calculations of nano-structured materials using diffraction data in combination with distance geometry methods is explained in [50]. Tools, results and perspectives of quantum software interfaced with crystal structure databases are given [51]. Density-functional theory electronic structure calculations of static and elastic properties and *ab initio* molecular dynamics simulations for poly-atomic systems made possible in package reported in [52] and first-principles computation of material properties using the ABINIT software package in [53]. A primer to efficient tight-binding molecular dynamics is given by Colombo [54]. CONQUEST code for large-scale *ab initio* calculations in materials science is presented in [55]. Prediction of materials properties by *ab initio* computer simulations is reported [56].

Atomistic Molecular Dynamic Simulations

Molecular Dynamics (MD) [57,58,59,60,61,62,63,64,65] is, in its simplest formulation, numerical integrations of Newton's second law (acceleration $(\mathbf{a}(t))$ of a particle is equal to the force $(\mathbf{F}(t))$ acting on a particle divided by the mass $(\mathbf{m})$ of the particle) applied on a collection of inter-



$$U(r_{ij}) = 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\ \varepsilon_0\ r_{ij}}$$

$$\sum_j^N \mathbf{F}_{i<j} = \sum_{j<i}^N \left(- \frac{\partial U_{ij}(r_{ij})}{\partial r_{ij}} \frac{\mathbf{r}_{ij}}{|\mathbf{r}_{ij}|} \right) = \mathbf{F}_i = \mathrm{m}_i \mathbf{a}_i = \mathrm{m}_i \frac{d\mathbf{v}_i}{dt} = \mathrm{m}_i \frac{d^2\mathbf{r}_i}{dt^2}$$

$$\mathbf{v}_i = \frac{d\mathbf{r}_i}{dt} \Rightarrow d\mathbf{r}_i = \mathbf{v}_i dt \qquad \mathbf{a}_i = \frac{d\mathbf{v}_i}{dt} \Rightarrow d\mathbf{v}_i = \mathbf{a}_i dt$$

$$\mathbf{r}_i(t+\Delta t) = \mathbf{r}_i(t) + \Delta t\, \mathbf{v}_i\,(t + \tfrac{1}{2}\Delta t) \qquad \mathbf{v}_i(t+\tfrac{1}{2}\Delta t) = \mathbf{v}_i(t - \tfrac{1}{2}\Delta t) + \Delta t\mathbf{a}_i(t)$$

$$\sum_{i=1}^N \frac{1}{2}\,\mathrm{m}\mathbf{v}_i^2 = \frac{3}{2}\,Nk_BT \quad \Rightarrow \quad \langle T \rangle = \frac{1}{M}\sum_n^M \left\langle \frac{1}{3Nk_B} \sum_{i=1}^N \mathrm{m}_i \mathbf{v}_{in}^2 \right\rangle$$

$$V = L * L * L \qquad \chi = \chi - L * \text{anint } (\text{x}* 1.0/\text{L})$$

Self-assembled Materials, Figure 1
Molecular Dynamics simulations in a nutshell

acting particles. Integrating the acceleration of each particle gives the velocity ($\mathbf{v}(t)$) and a subsequent integration of the velocity gives the actual position ($\mathbf{r}(t)$) of the particle. Thereby the trajectory of the particle can be determined. The forces momentarily acting on the particles (masses) are calculated as negative spatial derivatives of the potential functions $U(\mathbf{r})$ (force fields) describing the attractive and repulsive interactions between the particles (including the internal degrees of freedom): $\mathbf{F} = -\nabla U(\mathbf{r})$. Normally only the pairwise intermolecular interactions are included as an approximation in classical MD simulations of liquids and solutions. The number of particles (N) in current simulations is from thousands to hundreds of thousands and is a compromise set by the available computing power. Particles are inserted into a cell (box) with a volume V. The shape of the cell is preferably chosen to have translational symmetry so that replicas of the cell make a pseudo-infinite periodic system where particles are synchronously able to leave and enter the cells according to periodic boundary conditions (PBC). This keeps the particle density constant. So called minimum image convention connected to the PBC is applied to avoid artificial condensation effects (for more details see Table 1 and the text in connection to the figure). In Newtonian dynamics the energy (E) is expected to be conserved at equilibrium conditions and the simulation corresponds to a microcanonical (NVE) ensemble. By using suitable Hamiltonians containing thermostats canonical (NVT) ensembles can be constructed and with corresponding barostats isobaric (NPT) ensembles. Several other ensembles can be constructed for MD simulations as well.

Historically, Alder and Wainwright introduced the method of Molecular Dynamics 50 years ago by simulating a system of hard spheres [66,67] observing a phase transition, while Rahman [68] carried out the first simulation for a liquid system using argon atoms in 1964 giving fairly accurate diffusion coefficient. Verlet was able to obtain thermodynamic state functions noble gases [69]. A very important pioneering work for molecular systems was that of Rahman and Stillinger for liquid water in 1974 obtaining radial distribution functions in close agreement with experiments [70], which also started a real boom in computer simulations. MD simulation is a powerful method to study structural, thermodynamic and dynamical properties in gases, liquids and solids. Using MD the motion of molecules can be followed to gain a better understanding of chemical reactions, fluid flow, phase transformations, droplet formation, and other physical phenomena. MD is a statistical mechanical method (both equilibrium and non-equilibrium).

Figure 1 gives all the information needed to write a simple computer program to run MD simulations of N

charged particles in a cubic volume V. *First row*: Pairwise interaction potential consisting of Lennard–Jones (van der Waals) and Coulombic terms. Particle sizes are determined by the σ parameter and strength of the van der Waals attractions is given by ε. Particle charges are given by q in the Coulombic term and pairwise distances are given by r_{ij} (scalar) and the $\mathbf{r}_{ij}$ (vector). *Second row*: Forces acting on each particle are obtained as the negative spatial derivative of the potential function given in first row and are equal to particles mass times its acceleration (Newton's II law). *Third row*: By integrating the fluctuating time-dependent acceleration gives particle's velocity and a subsequent integration of the velocity gives the momentary position of the particle. *Fourth row*: Example of a numerical finite difference algorithm called the "Verlet leap-frog" to carry out the integration. Δt is the time step and is normally of order of a few femto seconds depending of the particle masses and the density of the system. Observe that the velocities in the Verlet leap-frog scheme are obtained as half time steps to increase the accuracy and stability of the algorithm. *Fifth row*: Example of how compute the average temperature from simulation data using the kinetic energy and the equipartition principle (each translational degree of freedom contributes 1/2 kT to the kinetic energy). M is the total number of time steps in the simulation. Other average quantities, for example the total energy (potential+kinetic energy) is calculated similarly. *Sixth row*: For a cubic simulation cell of a volume V (containing the N particles) the side length is L. Periodic boundary conditions (PBC) are easily applied on the positions (to keep the particles inside the cell) by the simple formula where the anint is a function which truncates the floating point number keeping only the integer digit. Example here is given for the x-components and should be done same way for y- and z-components. PBCs are applied on the distances as well according to the minimum distance convention to avoid artificial condensation effects during the simulations. Large scale simulations are carried out using parallel computers with a large number of processors. This requires computer software adapted to parallel architectures [71,72]. A large number of simulation packages are available from the authors developing them. Most of them are publicly available while some (most often those who are made particularly user-friendly by graphical interfaces) require a license to be purchased.

MD is an example of N-particle simulation techniques, a wide-spread method found virtually in all areas dealing with interacting dynamical particles from solvated electrons to protein-lipid interactions and from orientations of liquid crystals to formation of galaxies [58,59,60,61,62,63,64,65]. It is also the main simulation method in materials science from metals to zeolites and from polymers to colloids.

Self-assembled thiol monolayers on Au(111) made hydrophilic with polar end groups are studied with STM and MD simulations [73] to examine the role of the formation of hydrogen bonds between the molecules in the layer and with polar co-adsorbates (water and solvent). MD have been performed [74] to investigate the two-dimensional structure of organosilane self-assembled monolayers (SAMs). Unlike alkanethiol SAMs, the arrangement of molecules in organosilane SAMs is not crystalline. Simulations performed for structures with different bonding networks in the polysiloxane layer shows that the ratio of hydrogen bonds has a profound effect on conformations and strain energies. Water density profile close to spherical and planar hydrophobic objects is studied [75] using MD simulations. A substantial increase of the depletion layer thickness was found with the temperature. High electrostatic surface potential presumably plays an important role in the presence of charged solutes possibly promoting adsorption into the interfacial layer. MD investigations [76] of the morphology and structure of monolayer and multilayer of chiral molecule N-stearoy-L-glutamic acid (C-18-L-Glu) self-assembled on a mica surface show that hydrogen-bonding effects are the major driving forces in the layer formation proposing a multilayered model for the self-assembling. Static and dynamic properties of 2:1 layered silicates ion exchanged with alkyl-ammonium surfactants are studied using MD simulations [77] providing the structure and dynamics of the intercalated surfactant in agreement with experiments. Structure and properties of self-assembled monolayers (SAMs) of a bi-stable[2]rotaxane on Au(111) surfaces as a function of surface coverage are reported based on atomistic molecular dynamics (MD) studies with a force field optimized from DFT calculations with several experiments validating the predictions [78]. To develop strategies to self-assembly of nanostructured materials MD simulations for telechelic molecules composed of two polyhedral oligomeric silsesquioxane cages connected by one hydrocarbon backbone dissolved in liquid normal hexane were carried out [79]. MD simulations of nanoparticle self-assembly at a liquid-liquid interface are carried out showing in situ formation of clusters and migration of both single particles and clusters from the water phase to the trichloroethylene phase [80].

Among similar modelling works related to self-assembly could be mentioned recent advances of classical density functional theory with emphasis on applications to quantitative modeling of the phase and interfacial behavior of condensed fluids and soft materials, including colloids, polymer solutions, nanocomposites, liquid crystals, and biological systems are summarized [81]. Molecular mechanical methods are used to calculate strain distribution in self-assembled interfacial misfit dislocation arrays in highly mismatched III–V semiconductor materials in good agreement with experimental data using cross section high-resolution transmission electron micrograph images and also with other theoretical values [82]. Molecular self-assembly at equilibrium is studied using a family of simple directional and short-range van der Waals potentials giving rise to the self-assembly of linear polymeric, random surface, tubular, and hollow icosahedral structures. Dipolar potential with its continuous rotational symmetry about the dipolar axis contributes to chain formation, while higher multipoles lead to the self-assembly of open sheet, nanotube, and hollow icosahedral geometries [83]. To further understand the mechanism behind Layer-by Layer assembly processes requires even theoretical work, including thermodynamics calculations and molecular dynamics simulation as many kinds of physicochemical molecular interactions, including hydrogen bonding, charge transfer interactions, and stereocomplex formation are involved [84].

Monte Carlo Simulations

With modern computers the Monte Carlo technique has evolved from a classical numerical algorithm to solve multi-dimensional definite integrals based on statistical sampling to a very powerful simulation technique to study a large variety of complex systems with interacting particles [59]. The name “Monte Carlo”, given by Stanislaw Ulam [85], refers to the capitol of Monaco at Mediterranean, famous for its casinos and roulette tables. The corresponding “roulette” in the MC method is the random number generator supplying a sequence of numbers normally between 0 and 1 from a uniform random distribution (rand[0, 1]). MC methods can be used, for example, to compute equilibrium properties of classical many-particle systems. MC was in fact among the first applications run in the first computers in early 50s. The first liquid simulation using MC was carried out by Metropolis and coworkers [86]. As molecular computer simulation methods MC and MD are very similar until the particles start to move. The simulated systems in both MC and MD are built in the same way by inserting the N particles in a simulation cell of volume V, applying the periodic boundary conditions and choosing a force field to describe the interactions between the particles and the cut-off distances for short and long-range interactions. Ewald summation can be employed if found necessary. By fixing the temperature T we can perform the MC simulation in a canonical (NVT) ensemble which is also the simplest choice in molecular MC simulations. Recall that the NVE (microcanonical) ensemble is the natural choice in the case of Newtonian MD simulations. While the MD simulation method is completely deterministic in the sense that the particles always follow exactly same trajectories if same input is given, the Monte Carlo (MC) method is stochastic and the particles are moved as a random walk. How the particles are moved depends on the studied systems. In the case of small and rigid molecules (as a liquid) the molecules (center-of-mass) is translated (displaced) in x, y and z direction of a small distance one molecule at a time rotating it slightly around its principal axes. For larger and flexible molecules the individual atoms are slightly moved and in addition different types of specific movements are performed mimicking somehow the real motion of the molecule. For example MC simulations of polymers are undergoing residual “flips”, “rotations”, “crankshaft”, “slithering snake” moves and “end-bridging” plus many more. Probably the simplest MC technique to simulate interacting particles at a desired temperature is Metropolis Monte Carlo [86].

The mother of all Monte Carlo methods is the Markovian master equation:

$$\begin{aligned}&\frac{\mathrm{d}P(\vec{r}_i,t)}{\mathrm{d}t}\\&\quad=\sum_{\vec{r}_j} W(\vec{r}_j,\vec{r}_i)P(\vec{r}_j,t)-\sum_{\vec{r}_j} W(\vec{r}_i,\vec{r}_j)P(\vec{r}_i,t)\end{aligned} \tag{1}$$

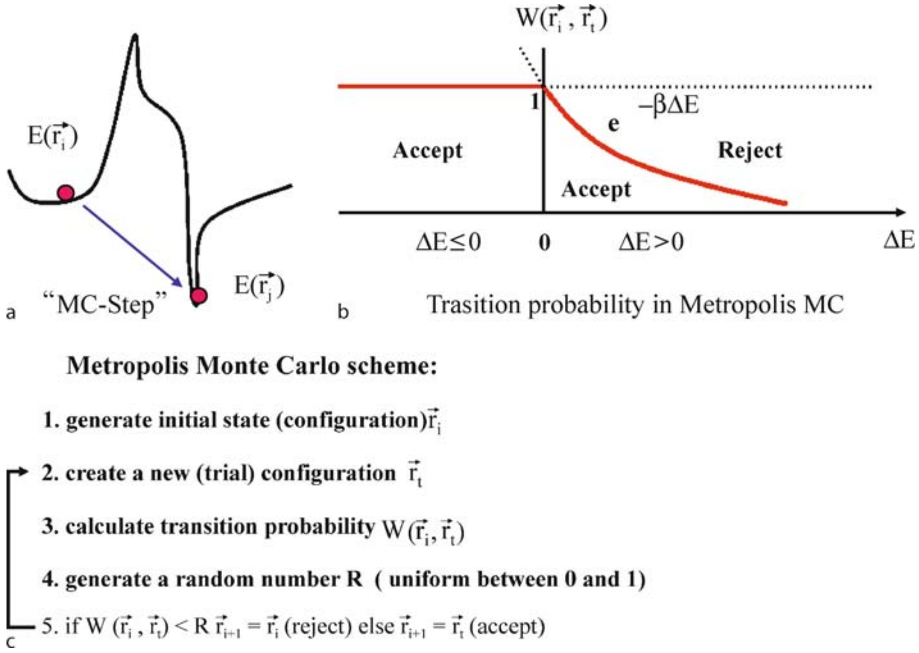
where $P(\vec{r},t)$ is the probability that the system is in state $\vec{r}$ at time t (time-dependent distribution of states/configurations) and $W(\vec{r}_i,\vec{r}_j)$ is the transition probability/rate per unit time for the system to undergo a transition from i-state to j-state. All MC techniques can be seen as methods for solving the master equation. At steady state (stationary or equilibrium):

$$\begin{aligned}&\frac{\mathrm{d}P(\vec{r}_i,t)}{\mathrm{d}t}=0;\quad t\to\infty\wedge P(\vec{r}_i,t)\to P(\vec{r}_i)\\&\quad\Rightarrow\sum_{\vec{r}_j} W(\vec{r}_i,\vec{r}_j)P(\vec{r}_i)=\sum_{\vec{r}_j} W(\vec{r}_j,\vec{r}_i)P(\vec{r}_j)\,.\end{aligned} \tag{2}$$

The detailed balance (microscopic reversibility) is valid. The canonical contribution is:

$$\sum_{\vec{r}_j} W(\vec{r}_i,\vec{r}_j)\mathrm{e}^{-\frac{E(\vec{r}_i)}{k_\mathrm{B}T}}=\sum_{\vec{r}_j} W(\vec{r}_j,\vec{r}_i)\mathrm{e}^{-\frac{E(\vec{r}_j)}{k_\mathrm{B}T}} \tag{3}$$

$$W(\vec{r}_i,\vec{r}_t)=\min\left(1,\mathrm{e}^{-\frac{E(\vec{r}_t)-E(\vec{r}_i)}{k_\mathrm{B}T}}\right)\,. \tag{4}$$



Metropolis Monte Carlo scheme:

1. generate initial state (configuration) $\vec{r}_i$
2. create a new (trial) configuration $\vec{r}_t$
3. calculate transition probability $W(\vec{r}_i, \vec{r}_t)$
4. generate a random number R (uniform between 0 and 1)
5. if $W(\vec{r}_i, \vec{r}_t) < R$ $\vec{r}_{i+1} = \vec{r}_i$ (reject) else $\vec{r}_{i+1} = \vec{r}_t$ (accept)

Self-assembled Materials, Figure 2

a Schematic illustration of a Monte Carlo move. b Transition probability *W* in Metropolis Monte Carlo. c Computational scheme for Metropolis Monte Carlo

However, it is not possible to calculate W from this equation. MC methods therefore use different functional forms for W. Transition probabilities in thermodynamical MC (like Metropolis) do not need to have *any* relationship to the dynamics (energy barriers) of the system. To assign real time to MC steps (Fig. 2a) is not possible. Conventional MC methods are only applied to sample systems in (or close to) thermal equilibrium. Functional form of transition probability in Metropolis MC:

In standard Metropolis MC the looping should not be done until accepted. All moves with probability W larger than 1 are accepted (the energy goes down) while if W is less or equal to 1 the move is accepted if the random number is less than the Boltzmann factor in Eq. (4) (see also Figs. 2). A rule of thumb is to choose the size of the displacements in such a way that the acceptance/rejection ration is close to 1/2 which might be considered as the upper limit while the lower limit is roughly 1/3 [87]. Several particles should not be moved simultaneously in Metropolis MC. There is no direct correspondence between a Monte Carlo step and real time but it is possible to use experimental information of for example diffusion to calibrate the MC simulations to obtain a rough estimate of real time.

Some examples of MC works in connection to self-assembly are for example the MC (both multi-scale NVT and grand canonical) simulations carried out to investigate nano patterns an self-assembly of surfactants inside SWCNT systems [88] and Monte Carlo simulations of gold nanocrystals and (111) slabs covered with alkyl thiols are carried out with and without explicit solvent (n-hexane) at $T = 300\,\mathrm{K}$. Inclusion of explicit solvent is found important in [89]. Self-organization and chain-forming of large ensembles of nanoparticles is studied by combining DLVO theory and MC simulations [90]. The growth of linear agglomerates is kinetically controlled by a high activation barrier from all of the directions except one at end of the chain. Adsorption of different model amphiphiles in apolar and polar solvents is investigated using MC simulations with a coarse grained model. As coating agents the surfactants with a single hydrocarbon tail or two branches are found to protect better the particle surfaces than amphiphiles with three or more branches [91]. Lattice Monte Carlo simulation is used to investigate the equilibrium between free surfactant molecules in aqueous solution and those adsorbed layers on structured solid. The solid surfaces are composed of hydrophilic and hydrophobic surface regions [92]. Lattice MC simulation of self-assembled ordered hybrid materials is reported and a comparison of structural characterization of the different phases using aggregate size distribution, density profiles, and radial distribution functions [192]. MC simulations are used to study self-assembly of symmetric diblock copolymers in confined state resulting novel self-assembled strip, circle, core-multishell, and multi-barrel layer structures [94]. Self-assembly and adsorption on hydrophobic surfaces are studied by MC methods in [95,96,97,98].

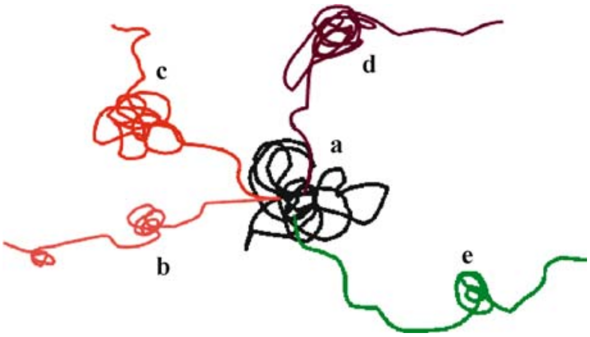
Kinetic Monte Carlo Simulations

Ideally we would like to have a general computer simulation method operational on atomistic level of matter which we can used for both short and long length and time scales. We wish to use it to simulate general dynamical processes with many different types of transitions between various states the system populates during its evolution towards a more permanent equilibrium state. The method of molecular dynamics (MD) is often the first choice. Using MD we model our system by applying realistic physical conditions. We also have correct real time in our simulation procedure. However, the time step Δt in atomistic MD simulation to numerically solve the equations of motion has to be chosen according to the frequency ($\omega_{\max}$) of the fastest dynamical event in the molecular system this being in practice $\Delta t \ll (\omega_{\max})^{-1}$. This makes atomistic MD simulations very inefficient if the events of interest occur beyond nanosecond time scales. Unfortunately a vast amount of motional modes and dynamical events take place beyond the capability of MD simulations even with the fastest computers around.

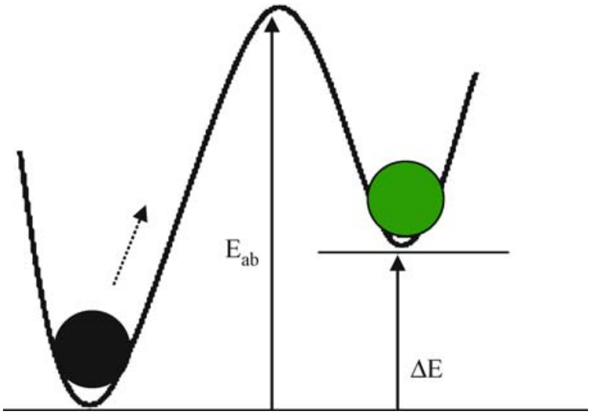
We can use Monte Carlo (MC), such as the common Metropolis method discussed above. We can build the molecular system similarly as in MD and impose the desired conditions onto it. We start simulations and pick up the particles randomly and one at a time moving them in small displacements based on the Boltzmann probability. Again, we realize that as the size of the system increases together with the number of motional degrees of freedom our simulation starts to take too long time. Besides, as mentioned in the previous chapter, we do not even know how to add the time into a Metropolis simulation as different events have different characteristic times for their dynamical behavior. MC as we perform it is normally only for sampling ensembles.

We now consider using a method called the kinetic Monte Carlo (kMC). Using kMC dynamical processes can be simulated with real time incorporated and all possible time scales can be covered easily. This sounds simply too good to be true so what is the “catch” here? Before going into details of kMC method (in fact there are many variants of it) we try to trace its origin by going back almost four decades ago and to computer simulations of vapor deposition on two-dimensional lattices [99] and crystal growth with surface diffusion [100]. These two works carry much of the basic underlying idea behind several kMC schemes developed later. The paper by Bortz, Kalos and Lebowitz (BKL) presents elegantly the theory and algorithms for a method finding immediately a strong position among physicists [101]. The BKL scheme is the solid framework in many kMC programs used today. In almost parallel and from purely chemical point of view Gillespie presented his stochastic simulation method for coupled chemical reactions [102] with the starting point being that on atomistic level the time evolution of chemical systems such as chemical reactions can be seen as both discrete and stochastic rather than continuous and deterministic as the text book kinetic equations are presented for us. Although the BKL method and the Gillespie method appear to have been developed in separate communities they both lead to same family of kMC methods [103]. The name kinetic Monte Carlo appears first as late as in 1992 [104].

Kinetic Monte Carlo is also based on solving the Markovian master Eq. (1) discussed in the previous Chap-



Self-assembled Materials, Figure 3
Accessible stationary states (b–d) from state a



Self-assembled Materials, Figure 4
Transition from state a to state b

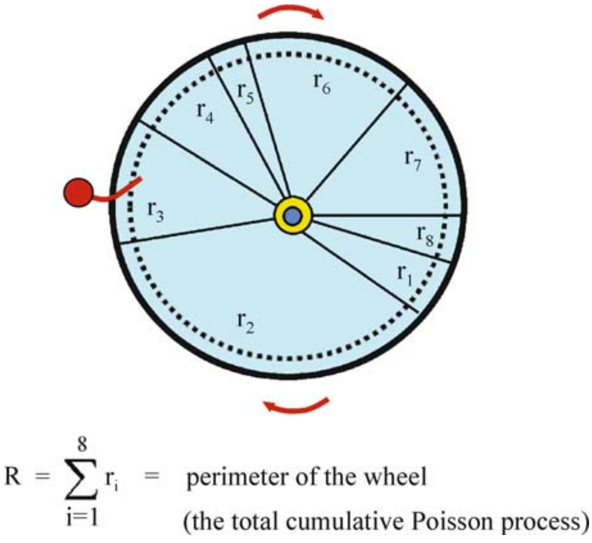
ter. In kMC we assume that the system is in a (stationary) state and that a number of other states are accessible from the current state (Fig. 4). A new state is randomly chosen among the neighboring states and a transition is made to this new state. The real time spent in the state before transition is added to the time development and can be seen as the time step in kMC. The time spent in a state (Fig. 4) is related to the transition rate:

$$r_{ab} = \upsilon \mathrm{e}^{-\frac{E_{ab}}{k_{\mathrm{B}}T}} \tag{5}$$

– where υ is the "attempt frequency". The time step is then calculated as:

$$\Delta t = -\frac{\ln(\mathrm{rand}[0,1])}{\sum\limits_{i=b,c,d,\dots} r_{ai}} \tag{6}$$

kMC requires that the simulated events are Poisson processes. There should also be a dynamic hierarchy among the processes. The real time increments (time steps) should be able to be calculated based on Eqs. (5)–(6). In practice the detailed balance (Eq. (2)) is not strictly required as the systems can be far from equilibrium. If the transition probabilities are: (i) independent on previous history, (ii) the same at all times and (iii) a uniform function of time, the the transitions are Poisson processes. A kMC transition should not have any effect on any other transition. Also the time when the next transition will occur should be independent from the occasion of the previous transition. To illustrate kMC we may consider a "wheel of fortune" (Fig. 5) where the sectors correspond to individual Poisson processes (events) and are not necessarily evenly divided. The perimeter of the wheel is the cumulative process (also a Poisson process). The corresponding dynamical hierarchy in Fig. 5 is as:


r6
r5
r4
r7
r3
r8
r1
r2
$R = \sum_{i=1}^{8} r_i$ = perimeter of the wheel
(the total cumulative Poisson process)

Self-assembled Materials, Figure 5
The Poisson "wheel of fortune" to illustrate the kMC

$$[0 \leqslant r_5/r_{\max} \leqslant r_8/r_{\max} \leqslant r_1/r_{\max} \leqslant \cdots \leqslant r_{\max}/r_{\max} \leqslant 1] \,.$$

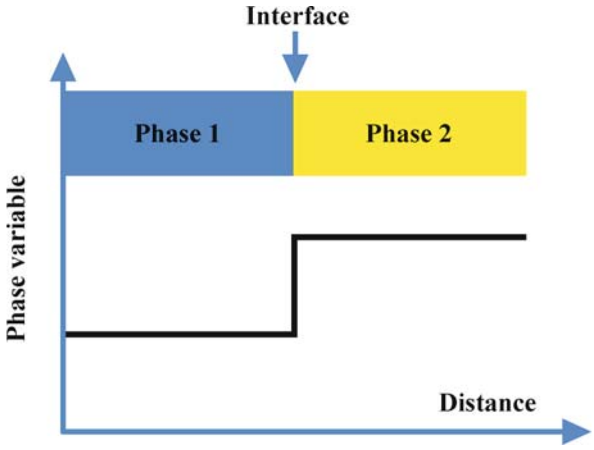
The whole spectrum of rates exists and the fastest rate (r_2 in Fig. 5) gives the highest probability. Note, the Metropolis MC does not fullfill dynamic hierarchy as all $\Delta E \leqslant 0$ transitions are given probability 1. After a kMC time step we are in a new state and we have a new "wheel" like the one in Fig. 5 but possibly with a different number of sectors corresponding to the new states with a finite probability to make transitions to. A new ensemble of processes is obtained after each transition. This also means that the time step in kMC changes continuously from one step to another. There are no restrictions how short or long the time step is: it can be for example a few femtoseconds or several hours. It simply follows the evolution of the system during the simulation. Time steps can be shorter in the beginning of simulation due to (initially) fast moving objects. Also, slow moving objects may break to smaller ones moving faster. Kinetic MC sounds like the optimal simulation method which can cover all possible time and length scales. The "catch" here is that the *user* has to supply both the states and the transition rates to the new states after each time steps. The kMC is not a self-going scheme in the same way as the other simulation methods are. This of course limits its use currently. Transitions rates can be obtained for example from quantum chemical calculations, MD simulations, transition state theory and experiments. As all the information of the system and rates is available the method has many attractive features: No thermodynamical equilibration is needed and no simulation of "dead dynamics" like in MD (event/step is always guaranteed). kMC is suitable for "driven systems" (irradiation, high pressure, etc.). kMC has been used in studies surface diffusion, crystal growth, molecular beam epitaxy growth, defect mobility, vacancy diffusion, reactions of surfaces, diffusion in zeolites, clustering dynamics, nucleic acid hybridization, protein folding, etc. The number of new application areas is growing rapidly. kMC can be easily incorporated in multiscale simulations schemes.

Kinetic MC is used to determine the critical layer thickness for misfit dislocations using deposition flux, temperature and a pairwise interaction potential between the particles as parameters [105]. Kinetic MC simulations are used to study the deposition rate dependence of nanopattern formation on periodically strained surfaces. The optimum nanopattern quality depends on surface strain field, temperature and deposition rate and the amount of material deposited [106]. Self-assembly of the elementary building blocks of nanophase materials are studied using both a kinetic mean-field model and a mesoscopic approach in which self-assembly is viewed as an encounter-controlled process on a discrete lattice [107]. A 3D kMC model with a Green's function-based long-range strain energy contribution and with an up-down ratio for atoms to jump out of the plane of the surface is developed to simulate the growth of self-assembled quantum dot islands [108].

Phase-Field Modeling

Classical (macroscopic) theories & models assume sharp interfaces (infinitely thin with a specific surface tension) between phases and thermodynamical variables such as T, P, concentration, etc. are used independently in each phase obeying the phenomenological rules (schematically shown in Fig. 6). Phase diagrams are common tools in chemistry text books and traditionally the compositional and structural evolution of the phase regions have been treated mathematically using distinct interfaces with suitable boundary conditions at the interfaces. The idea of a *sharp* interface is old going back to early 19th century and Young, Laplace and Gauss. Gibbs later presented the interface as a dividing surface so that the properties were distinctly different on both sides of the interface (Fig. 7). Subtraction of the upper curve (real system) from the idealized system below gives the corresponding excess quantity (here concentration) and sharp mathematical surface line. The Gibbs dividing surface has zero volume but non-zero excess quantities.

Also the idea of a *diffuse* continuous interface (Fig. 8) goes back to 19th century when it was thought that gas and liquid states were simply distant points of the same condition of matter and it was possible to move from one phase to another along a continuous path. This picture was established both experimentally by Andrews and theoretically by Thompson and van der Waals [109]. Compared to singularly sharp interface model there are many benefits with using a diffuse inteface with a finite interface volume. All field variables and corresponding equations can be defined simultaneously for the whole system. Diffuse interface allows modelling of various types of physical variables, systems and effects in a flexible and efficient way. Diffuse inteface exists on atomic scale as the surface layers are rough on atomic scale rather than mono-molecular. The density profile at the interface becomes larger than the atomic length scale due to the roughness.



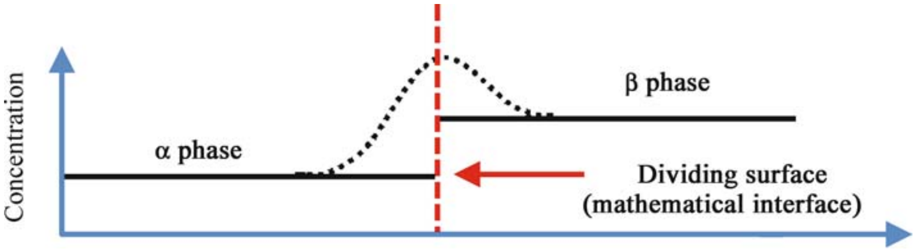
Self-assembled Materials, Figure 6
Sharp limit between two phases

Phase field modelling gathers all categories of diffuse interface models to describe a wide variety of materials phenomena (for excellent reviews, see [110,111,112,113, 114,115,116,117]). It can describe easily a large number of phenomena from nucleation to solute trapping as different physical effects can be easily included [115]. It is based on two assumptions: (i) There exists a continuous phase field variable $\phi(x, t)$ which characterizes the phases and interfaces of the system at any point point in space and time. In bulk phases it is given a constant value. (ii) The total (Helmholtz) free energy, $F[\phi, \ldots]$ can be given as a functional of the phase field variable ϕ together with any other thermodynamical variable, for example temperature, concentration etc. Even other variables can be introduced depending on the studied system. A general form of F is given as [113]:

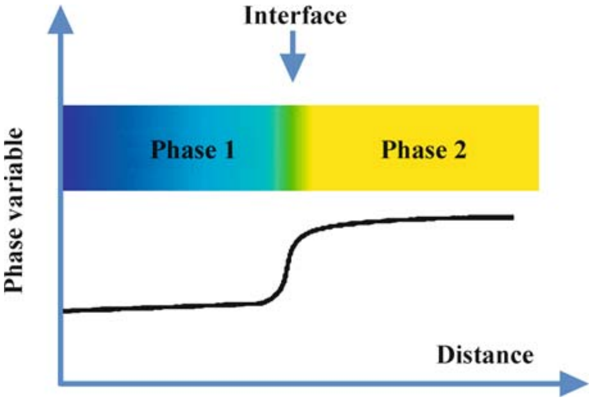
$$F = \int \Big[f(c_1, c_2, \ldots, c_n, \eta_1, \eta_2, \ldots, \eta_p) + \sum_{i=1}^{n} \alpha_i (\nabla c_i)^2 + \sum_{i=1}^{3} \sum_{j=1}^{3} \sum_{k=1}^{p} \beta_{ij} \nabla_i \eta_k \nabla_j \eta_k \Big] \mathrm{d}^3 r + \iint G(r - r') \mathrm{d}^3 r \mathrm{d}^3 r' \tag{7}$$

c denotes *conserved* variables (satisfying local conservation conditions)
η denotes *non-conserved* variables.
f is local free-energy density,
α and β are coefficients for energy gradients.

The first integral in Eq. (7) contains the local short-range interactions to the free energy, while the second integral represents the non-local long-range interactions, such as, elastic, electrostatic, polar, etc. Various phase-field models differ in their ways to treat all these contributions



Self-assembled Materials, Figure 7
Gibbs dividing surface



Self-assembled Materials, Figure 8
Diffuse continuous interface between two phases

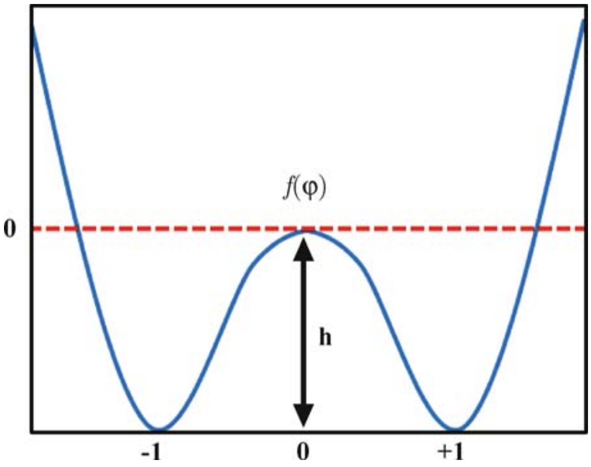
to the total free energy functional. The local free energy f is the most important term in the phase field models:

$$f = f(c_1, c_2, \ldots, c_n, \eta_1, \eta_2, \ldots, \eta_p)\,.$$

A large variety of these functions have been proposed. In many phase-field models (like for solidification) simple double-well potential functions are used for local free density like the one given in Eq. (8) and in Fig. 9:

$$f(\varphi) = 4h\left(-\tfrac{1}{2}\varphi^2 + \tfrac{1}{4}\varphi^4\right)\,. \tag{8}$$

The variable ϕ in Eq. (8) is normally called order parameter. It varies smoothly from one phase to another, distributing interfacial forces and other interactions over the interface region [8]. We can often use it in a simple double-well potential for a case where the values $\phi = -1$ and $+1$ represent liquid and solid state, respectively, h is the barrier height between the two states. We may also have the phase field $\phi = -1$ and $+1$ representing two different heterogenous mixtures during an isostructural (spinodal) decomposition while 0 is a single homogeneous phase and h is the driving force for the transformation. Other types of



Self-assembled Materials, Figure 9
Simple double-well free energy density function

similar simple local free energy density functions can also be found in literature.

Some other forms of local free energy profiles are for example the "double-obstacle" potential:

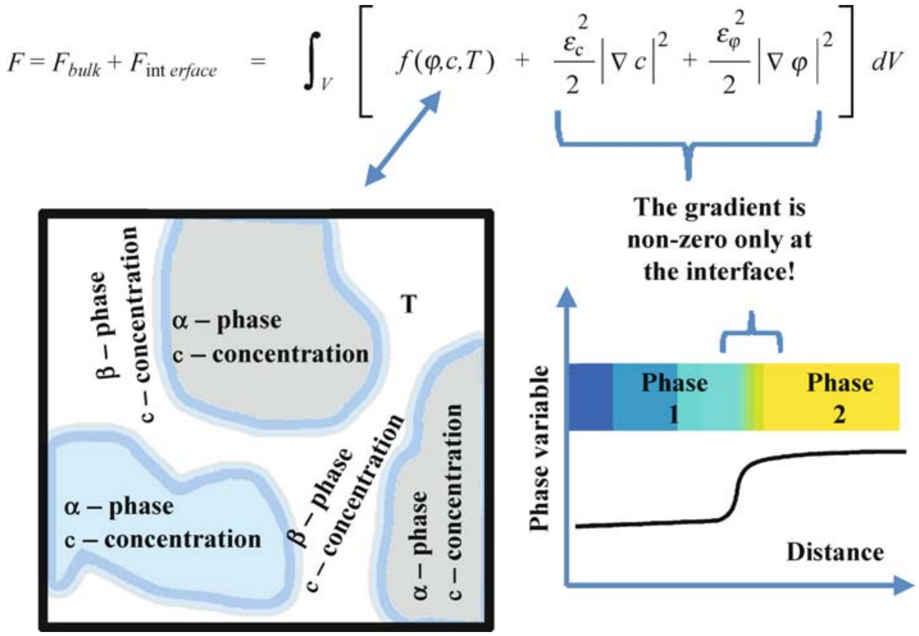
$$f(\varphi) = h(1-\varphi^2) + I(\varphi)$$
$$I(\varphi) = \begin{cases} \infty & \text{if } |\varphi| > 1 \\ 0 & \text{if } |\varphi| \leqslant 1\,. \end{cases} \tag{9}$$

Or a "crystalline" potential having infinite number of minima:

$$f(\varphi) = h\sin(\pi\varphi)\,. \tag{10}$$

Often additional phase field parameters like temperature are needed, T = temperature, T_m melting point and α is a positive constant:

$$f(\varphi, T) = 4h\left(-\tfrac{1}{2}\varphi^2 + \tfrac{1}{4}\varphi^4\right) + \frac{15\alpha}{8}\left(\varphi - \tfrac{2}{3}\varphi^2 + \tfrac{1}{5}\varphi^4\right)(T - T_\mathrm{m})\,. \tag{11}$$



Self-assembled Materials, Figure 10
Two main contributions to free energy: bulk and interface

Often coupling of phase field variables are needed – like in the following function to describe the spatial distribution for grain growth (for grains with different orientations):

$$f(\varphi_1, \varphi_2, \dots) = 4h\left(-\tfrac{1}{2}\sum_i \varphi_i^2 + \tfrac{1}{4}\sum_i \varphi_i^4\right) + \alpha \sum_i \sum_{j>i} \varphi_i^2 \varphi_j^2 \,. \qquad (12)$$

When $\alpha > 2h$ (in Eq. (12)), infinite number of of minima are located at $(1, 0, \dots)$, $(0, 1, \dots)$, $(-1, 0, \dots)$, etc., representing possible orientations of grains in a polycrystal. Functions in Eqs. (9)–(12) are discussed in [113] where several other types local free density functions are found. In general there are a large number of free energy functions like those above, all designed for specific applications and studies of phenomena. For many studies like solid-state transformations the energy functions are made of well-defined physical order parameters and use required symmetry operations of the studied high temperature phases. In general, the interfacial energies should be seen as anisotropic due to the crystalline feature of solid phase [113]. Multiple order parameter models have been proposed using both thermodynamical and geometrical formulations [114].

There are two main contributions to the total free energy; bulk and interface as illustrated in the example in Fig. 10.

Consider the isothermal process from the example in Fig. 10 with a decreasing total free energy F. At equilibrium, if the gradient coefficients are constants, the variational derivatives of F satisfy following equations [114]:

$$\frac{\delta F}{\delta \varphi} = \frac{\delta f}{\delta \varphi} - \varepsilon_\varphi^2 \nabla^2 \varphi = 0 \qquad (13)$$

$$\frac{\delta F}{\delta c} = \frac{\delta f}{\delta c} - \varepsilon_c^2 \nabla^2 c = \text{constant} \,. \qquad (14)$$

The Eq. (13) is for non-conservative variables while Eq. (14) is for conservative variables. Equation (14) is constant as the amount of solute is constant in the volume (concentration is one of the conserved quantities).

Time evolution of the phase field is assumed to be proportional to variation of the total free energy functional with respect to the order parameter ϕ [110]: $\partial\varphi/\partial t \;\propto\; L(\delta F/\delta\varphi)$. Where L = partial differential operator $(L(0) = 0)$. The phase field equations are given:

Cahn–Hillard equation [118,119]:

$$\frac{\partial c}{\partial t} = \nabla M \nabla \frac{\delta F}{\delta c} = \nabla M \nabla \left(\frac{\delta f}{\delta c} - \varepsilon_c^2 \nabla^2 c\right) \qquad (15)$$

– for conserved variables such as: concentration, molar fraction, mass, etc.

Allen–Cahn equation [120,121]:

$$\frac{\partial \eta}{\partial t} = -L\frac{\delta F}{\delta \eta} = -L\left(\frac{\delta f}{\delta \eta} - \varepsilon_\eta^2 \nabla^2 \eta\right) \qquad (16)$$

– for non-conserved variables such as: order parameters, phase fields, etc. This equation is also known as the time-dependent Ginzburg–Landau equation.

The evaluation of the phase field variables can be obtained by solving the kinetic Eqs. (15) and (16) above. Finite difference methods with explicit time-stepping are normally used by constructing a uniform grid. Discussion of numerical methods and stability of the computational schemes are given in [122].

The areas of applications of phase field simulations are increasing rapidly. Solidification and solid-state transformations are examples of the use phase field modelling as well as coarsening, grain growth and dislocation dynamics. A list of several applications can be found in Table 1 of [113].

A further development of phase field theory, called phase-field crystal (PFC) method, was present by Elder and Grant [123] for elastic and plastic deformations, free surfaces and multiple crystal orientations.

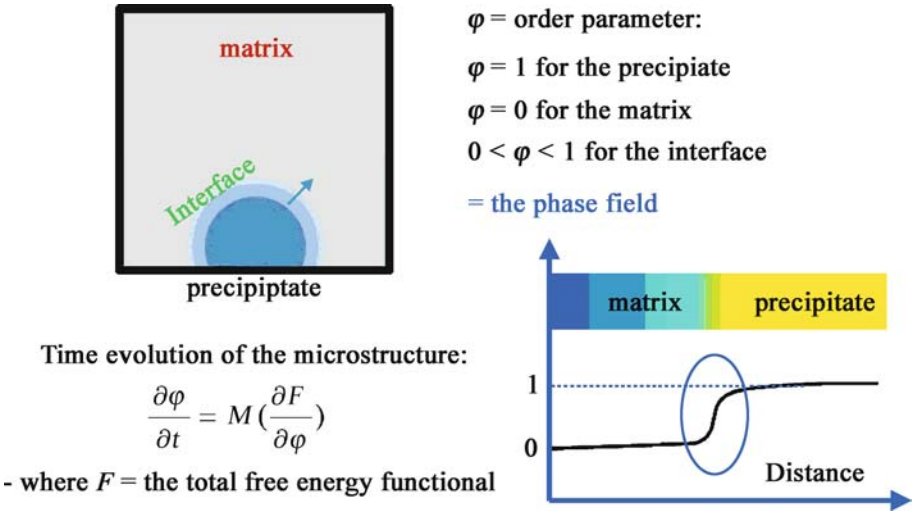
Continuous phase field model by Lu and Suo have been found applicable on self-assembly and monolayer growth on surfaces describing all required components: the phase separation, phase coarsening and phase refining processes [112]. Lu and Salac [124] have developed a phase field model to simulate pattern formation due to electric dipole interactions among adsorbate molecules revealing a unique self-assembly behavior. Wang [125] has developed a diffuse interface field approach (DIFA) which is capable to explicitly describe short- and long-range interactions and treat arbitrarily shaped particle packing processes and other related applications.

Molecular dynamics simulations can be combined with phase field modelling. Two basic approaches have been used where MD simulations carried out at the interface region are connected the phase field simulations using finite element methods [126,127] or where (moving region) MD simulations are carried out at the interface region to supply parameters, such as the strain response, kinetic coefficient (velocity vs. under-cooling behavior) and diffusivity, to the phase field model [128]. As the microstructure in many phase-field models is assumed to be coherently uniform, which is, in many cases incorrect, atomistic MD simulations may be of help in supplying anisotropic information about structure and dynamics of interface regions. A very interesting theoretical model for binary crystal nucleation was developed by Gránásy and coworkers [129,130,131,132,133,134,135,136,137,138,139], and incorporated into simulations and applied on Lennard–Jones and ice-water systems [129,130,131], polymeric dendrites [132,134,136], hydrate formation by Gránásy, Kvamme and coworkers [133,135,137,138] and heterogeneous crystal nucleation [139]. Davidchack and Laird [140,141] have used MD simulations with hard spheres to study interfacial planes between crystal and fluid-like domains.

Plapp and Karma have proposed a robust hybrid multi-scale scheme applicable in one, two and three dimensions by combining diffusion Monte Carlo (similar to the one used to solve Schrödinger equation) and phase-field models to perform simulations of dendrictic solidification. The method is applicable for other applications than solidification [142]. General critical discussion of microstructure modelling is given in [143] highlighting phase field and cellular automata methods. More examples of multi-scale methods will be given below in a dedicated chapter.

Random Sequential Adsorption

Random sequential adsorption (RSA) is a family of models for a variety of phenomena in chemistry, physics, biology, ecology, sociology and in many other fields as it is a general and highly interesting mathematical and physical packing problem [144]. It is applicable on all kinds of moving particles from atoms to motor vehicles arriving randomly as a Poisson process selecting a free space based on a distribution of sites or slots. Particles become bound if they find a vacant place (with no overlap with other particles already occupying the actual space), or they leave the system in case of no luck (possibly to try again). The process continues for a certain period of time until the distribution of sites becomes saturated (complete packing) following the random packing scheme, or if no more particles are arriving although there may still be sites available. For excellent reviews see [145,146,147,148,149]. Colloids and Surfaces A, vol. 165 is a special issue containing several articles of RSA. The one-dimensional RSA model is known as the "car parking" problem as it can be seen as an illustrative and familiar analog. Imagine a long street with apartment houses where people and families live. During the day (office hours) there is plenty of space to park but in the evening the parking spots become quickly occupied (adsorption). In the next morning cars one after another are driven away (desorption). The solution of this 1D RSA problem would give us an answer of how many drivers on average (driving similar cars) can park at random positions along a street of a certain length? This can be solved analytically if the cars are equally long and the parking slots are discrete. For continuous space along the street the exact solution was given by a Hungarian mathematician Alfréd Rényi [150]. Higher dimensional problems are complicated and approximate solutions are given


matrix
Interface
precipiptate
φ = order parameter:
φ = 1 for the precipiate
φ = 0 for the matrix
0 < φ < 1 for the interface
= the phase field
matrix
precipitate
1
0
Distance
Time evolution of the microstructure:
$\frac{\partial \varphi}{\partial t} = M\left(\frac{\partial F}{\partial \varphi}\right)$
- where F = the total free energy functional

Self-assembled Materials, Figure 11
A phase field model for a growth of a precipitate

based on series expansion, or as a sequential Markovian process solved using Monte Carlo simulations although the detailed balance does not apply for irreversible processes.

The two dimensional RSA is an important and reliable model for molecules adsorbing on a surface reversible or binding irreversibly (with covalent bonds) to specific sites on the surface. Molecules can typically be monomers, dimers (occupying two neighboring sites which are vacant). Polymers can be adsorbed if they find successive empty slots on the surface. Various rules can be imposed on gas molecules approaching already occupied site. The interactions can be both specific and non-specific depending on the studied systems. Orchestration of van der Waals, electrostatic, hydrophilic/phobic interactions as well as hydrogen bonds covalent and ionic bonds leads to adsorption of molecules onto surfaces (consisting of atoms and molecules). Collective sequential adsorption (CSA) is a further generalization of RSA taking into account the local environment around the adsorption site [146]. Even large biomolecular systems (e. g. proteins, DNA, colloids, cells) can be coarse-grained and treated with RSA methods. RSA have been used even for translocation of biopolymers through pores [151] and nuclesomes on a stretched single strand DNA [152]. Generalized RSA methods can used to simulate multilayer buildup in a self-assembled layer-by-layer (LbL) process [153]. Good agreement was obtained between the simulations and experiment for colloids and poly-electrolytes suggesting that the method could be extended to polymers and proteins. Kinetics, jamming limit and structural phase behavior of polydisperse tethered nanoparticles are studied by Gray et al. [154]. Dynamics of self-assembled polyelectrolyte multilayers on glass substrates have been studied by Breit and coworkers [155]. The kinetics could be quantitatively understood using a RSA model for the buildup of a film consisting of polyelectrolyte disks with polydisperse sizes. RSA types of methods in modelling of self-assembled monolayers of charged colloidal particles are reviewed in [156]. Erban and Chapman have presented a diffusion driven random sequential adsorption simulation model operating in real physical time containing features for both reaction kinetics between the molecules and the surface and geometrical constraints of the molecular surface. It is illustrated by an assembly of reactive polymers on a virus surface [157]. Pre-patterning is used as a tool in RSA-based lattice Monte Carlo simulations to improve the self-assembly in nano and micro-scale structure engineering by Cadilhe et al. [158]. RSA model is applied for gelatin self-assembly in binary mixture of water and ethanol [159]. Short-time and long-time kinetics in colloidal adsorption to a monolayer are studied by combining RSA model and Brownian dynamics by Gray and Bonnegaze [160] in the case of self-assembly to meso-structures.

Cellular Automata

The Cellular Automaton (CA) concept was created by John von Neuman and Stanislaw Ulam more than a half century ago to originally solve problems in evolutionary biology [161]. CAs are discrete dynamical systems where space and time variables and all properties have finite discrete values. CAs start from simple identical (or nearly identical) individual systems interacting locally but evolve to complex systems following simple rules in a synchronous manner. CAs are similar to Petri nets [162] in providing a less mathematical way to differential equations and calculus. Other similar automata are the Turing systems, presented by Alan Turing in 1952 [163] showing how a simple mathematical model can describe spontaneously spreading of reacting chemical specie giving stationary concentration patterns. Turing patterns can also be found as stripes on zebras and spots on cheetahs and elsewhere in nature. Maybe the most widely known application of CAs is the tic-tac-toe like "Game of Life" by John Conway from 1970, a very simple model simulating birth and death of cells in a square lattice interacting according to a set of Boolean conditions. It was easily adapted from paper sheets to computer screen. Dynamic Cellular Automata (DCA) by Wishart and coworkers [164] is a simulation approach allowing a Brownian type of stochastic motion of individual molecules. DCA is more flexible as it allows several cells/molecules to move in a single step. Lattice gas automata are another common use of CA which together with the Boltzmann equation and kinetic theory has led to development of Lattice Boltzmann simulations [165], a powerful method to do fluid dynamics and to solve the Navier–Stokes equations in a from-bottom-up way. The obvious drawback in the current CA methods is that the objects do not have realistic features.

The quantum-dot cellular automata (QCA) approach is proposed as an alternative to molecular electronics. Theoretical behavior of QCA (arrays of Coulomb- coupled quantum-dot cells) is studied in [166] and the state of each QCA cell is determined by its interaction with neighboring cells through the Coulomb interaction and demonstrated experimentally [167] and the field was reviewed [168,169, 170]. A single-molecule implementation of a QCA cell is presented combined with *ab initio* electronic structure calculations [171] using simple prototype molecular systems with a molecule in which charge is localized on specific sites and can tunnel between those sites while the role of the dots is played by redox sites, with tunneling paths provided by bridging ligands. Possibilities to coarse-grain cellular automata, emergence, and the predictability of complex systems are studied by Israeli and Goldenfeld [172]. As molecular self-assembly is driven by local, short-range forces and therefore the dynamics is solely based on local interactions and as atomistic simulations of self-assembly become quickly complicated, a cellular automata based simulation, in which data structures, representing different molecular entities such as water and hydrophilic and hydrophobic monomers, share locally propagated force information on a hexagonal, two-dimensional lattice is proposed, with the purpose of this level of description to gain insight about entropy-driven processes in molecular many-particle systems [173].

Nilsson and Rasmussen present a lattice gas technique for simulating molecular self-assembly of amphiphilic polymers in aqueous environments, where water molecules, hydrocarbons tail-groups and amphiphilic head-groups are explicitly represented on a three dimensional discrete lattice [174]. A Cellular Automata Model of Diffusion in Aqueous Systems show that lipophilic solutes diffuse faster than do polar solutes (in agreement with experiments) [175]. Cellular automaton and MC simulations for the case of CO oxidation on a catalytic surface with simultaneous adsorption, reaction, and diffusion, including the Eley–Rideal step in the reaction mechanism is presented [176]. As properties of diffusing species in microporous materials are strongly influenced by the confining framework, providing the energy landscape for the transport process. Because of the simple topology and the cellular nature of the cages of a zeolite Demontis et al. [177] suggest that it is appropriate to apply to the study of the problem of diffusion in tight confinement a time-space discrete model such as a lattice-gas cellular automaton.

Multi-Scale Modeling

Molecules are moving in a world where "nanometer" could be used as a standard length unit and "femtosecond" is a convenient time unit. For us humans "meter" and "second" are the familiar units to which we all can relate our daily activities. In measuring the time from the big bang and distances between galaxies, both meter and second turn out immediately not to be very convenient to use. Obviously phenomena occur at different length and time scales. Connecting time and length scales has rapidly become a very vital area in from-bottom-up strategies. It will certainly still take quite some time for multi-scale modelling of matter to mature. Below we give examples of both concurrent and hierarchical approaches to multiscale modelling, the two main approaches.

"Concurrent" multiscale modelling can be accomplished by making geometric network of grid points describing the system at a detailed level successively coarser while losing more and more details. For example a crystal could be described with atoms or molecules in crystal sites connected together in 3D net. By applying a scheme with successive coarsening of connecting grid points approaching a continuum description can be coupled to the atomistic scheme by finite element methods [178]. A specific hand-shake region is created between the both scales. This is used to model cracks and fractures both at atomistic and macroscopic scale.

"Hierarchical" multiscale modelling can be performed by successively reducing the degrees of freedom which are not needed in the next less accurate description of the studied system. This makes the modelling faster and allows longer length and time scales to be covered. A powerful method called the Inverse Monte Carlo (IMC) was presented by us [179] where the "Inverse Problem" was solved completely to obtain interaction potentials (force fields) from radial distribution functions (RDF). The IMC constructs effective potentials consistent with the original RDFs. It can be used for example by starting from *ab initio* MD simulations and construct atomistic interaction potentials to be used in classical all-atom simulations. In all-atom simulations IMC can construct coarse-grained potentials for example by grouping together functional groups in large molecular systems and eliminate the solvent by incorporating it into the coarse-grained potentials. These coarse-grained potentials reproduce the RDFs in the all-atom simulations carried out with explicit solvent. This scheme can be used even beyond this kind of meso-scale description. There are other related methods presented [180,181].

Innovative computational schemes for multiscale modelling are reported continuously based on varying strategies. Examples below are given of studies (i) to couple first-principles simulations and other quantum schemes (ii) to use results from atomistic simulations to carry out subsequent coarse-graining (iii) to connect atomistic schemes with continuum models (iv) to apply kinetic Monte Carlo and other kinetic schemes subsequent to particle simulations, as well as (v) other combinations of various methods discussed earlier in this article are listed below, including a beautiful text-book example (vi) of multi-scale modelling of human hair.

(i) Transparent interface is created between classical molecular dynamics and first-principles molecular dynamics combining first-principles Born–Oppenheimer local spin density molecular dynamics (BO-LSD-MD) with classical molecular dynamics in [182]. Effective quantum mechanical classical mechanical (QM/CM) partitioning method for multiscale modeling is proposed in [183]. Quantum mechanics at the core of multi-scale simulations with neglect of diatomic differential overlap theory (used in common semi-empirical schemes such as AM1, PM3 and MNDO) is considered in [184]. Potential parameterization based on extended model systems where the force data is calculated from QM methods on a limited range of applications is shown to be essential for a consistent and ultimately, predictive embedding approach to concurrent multi-scale materials simulation in [185].

(ii) Automatic coarse-graining of polymers by deriving effective potentials for multi-atom units or super-atoms from atomistic simulations is reported in [186]. Hierarchical multi-scale modelling of plasticity of submicron thin metal films allowing modelling of thicker films a discrete dislocation model of "diffusional creep" is presented in [187]. Materials modeling platform bridging the molecular characteristics of polymers with macroscopic properties is given in [188]. Linkage between atomistic and meso-scale coarse-grained simulation based on model where the force acting on the CG particles is divided into the mean force (calculated from constrained MD simulations), friction force and random force is discussed in [189,190]. Combined atomistic and meso-scale simulation of grain growth in nanocrystalline thin films with input materials parameters obtained by MD simulation in [191]. Fuzzy clustering approach to hierarchical MD simulation of multiscale materials phenomena by combining a hierarchy of sub-dynamics: (i) rigid-body cluster dynamics for global conformational changes; (ii) implicit integration of Newton's equations for the coalescence of the clusters; and (iii) normal-mode analysis of fast atomic oscillations is used to facilitate the seamless integration of the multiple levels of abstraction in [192]. A multi-scale simulation of tungsten film delamination from silicon substrate where MD simulations of a single crystal *W* block under tension are used to calibrate a new decohesion model to investigate the effect of specimen size and loading rate on the material properties in [193].

(iii) Coupling of atomistic and continuum models in computational materials science using finite element methods is studied in [194] and finite-element method multi-scale atomistic-continuum modelling of crack propagation in a two-dimensional macroscopic plate by coupling the crack dynamics at the macro scales and nano-scales via an intermediate meso-scale continuum is carried out with molecular dynamics simulation driving the crack tip forward in [195].

(iv) Molecular simulations in zeolitic process design using configurational-biased Monte Carlo simulations combined with transition rate theories kinetic Monte Carlo and molecular dynamics simulations [196].

Multi-scale modeling of hydrogen isotope diffusion in graphite is carried out where molecular dynamics calculations resolve the microscopic length scale and deliver reliable input data for kinetic Monte Carlo calculations (jump frequencies, migration energies, jump step-sizes) from meso-scale up to the macroscopic system length [197]. An MD study of the carbon-catalyst interaction energy for multi-scale modelling of single wall carbon nano-tube growth is presented [198]. A multi-scale approach to radiation-induced segregation at various grain boundaries based on a new rate equation model and MD simulations is reported in [199]. Multi-Scale Computational Framework by integrating a Computational Fluid Dynamics software, a Kinetic Monte Carlo solver and an MD simulator for the self-assembly of atoms into molecular structures is presented in [200]. A novel polycrystalline thin film growth simulator with an atomic level one-dimensional kinetic lattice Monte Carlo model and a real time feature scale two-dimensional facet nucleation and growth model is demonstrated in [201]. A multi-scale atomistic study of the interstitials agglomeration in crystals using a hierarchy of atomistic approaches: the tight-binding molecular dynamics, molecular dynamics based on environment dependent inter-atomic potentials and a lattice kinetic Monte Carlo (LKMC) I reported in [202]. Computer simulations, spanning across different time and length scales, are used to study thin film growth morphology in organic self-assembled monolayers using thiophenes on gold. *Ab initio* calculations created a catalog of the energetics in vacuum and interactions in three orthogonal orientations to a Au (111) surface in [203]. This information was supplied as the input for kMC simulations of dimer and trimer representations of small organic molecules to describe both sub-monolayer and multilayer growth. Finally, MD studies were used to understand the packing structures of stable polymorphs of thiophene SAMs.

(v) Studies of behavior and structure of charged surfaces on different length scales is carried out using a different computational schemes yielding different levels of description of charged surfaces. (a) Coarse grained MC simulations of idealized surfaces incorporate large-length-scale fluctuation and correlation effects in the counter-ion cloud at a charged surface. (b) Brownian dynamics simulations of more realistic and structured surfaces give modified counter-ion distributions, also allowing estimations of mobility of counter-ions at charged surfaces. (c) All-atomistic MD simulations reproduce water structuring effects at surfaces such as hydration and hydrophobic de-wetting. (d) *Ab initio* calculations finally give the effective interactions between oppositely charged groups in vacuum and in solution in [204]. Multiscale simulations using generalized interpolation material point method and Structured Adaptive Mesh Refinement Application Infrastructure parallel processing as a multiple length scale tool from nanometer to millimeter is presented in [205]. To simulate the agglomeration carbonaceous nanoparticle assembly was studied using a multiscale coarse-graining by starting with an atomistic ensemble of 10 000 nanoparticles (or effectively 2 million total carbon atoms). The coarse-graining was accomplished applying a force-matching procedure. The results show rich and varied clustering behaviors for different particle morphologies [206]. A multiscale coarse-graining to derive coarse-grained models is applied to C60 and to carbonaceous nano-particles produced in combustion environments. The coarse-graining of the inter-particle force field is accomplished applying a force-matching procedure from all-atom MD simulations reproduce accurately the structural properties of the nano-particle systems [207].

(vi) A very impressive multiscale modelling approach to the mechanics of human hair fibres is given by Akkermans in [208] providing a very concrete example of multiscale modelling work-flows at their best where mesoscale models are constructed from atomistic simulations and where meso-scale simulation methods are used as input to finite-element calculations.

There are several excellent reviews discussing a great variety of invented multi-scale modelling schemes in materials science listed below and well worth of consulting. Simulation methods including broad areas of quantum mechanics, molecular dynamics and multiple-scale approaches, based on coupling the atomistic and continuum models are discussed in [209] and continuum/quasi-continuum approaches, the kinetic Monte Carlo technique and accelerated molecular dynamics simulation are gathered in [210]. Recent advances in bridging scale between quantum mechanical and continuum coupling are briefly described in [211] and multiscale hybrid simulation methods for material systems based on tight-binding DFT, MD and continuum models for deformation and diffusion on surfaces are reviewed in [212]. A seamless coupling of quantum to statistical to continuum mechanics involving models for unifying finite elements, molecular dynamics and semi-empirical tight-binding and coarse-grained Molecular Dynamics as an effective model in solid state is explained in [213]. Current trends from atomistic simulation towards multiscale modelling of materials based on quantum mechanical, especially density functional theory for electronic properties linked to atomic/molecular dynamics and kinetic Monte Carlo simulations where coarse graining leads to lattice gas and cellular automata, and also to continuum equations solved by finite-element and finite-difference techniques are discussed in [214]. Vvedensky presents a comprehensive outlook and an excellent review of current multiscale modelling methods for nanostructures across a vast range of length and time scales clearly stating that a complete understanding of the behavior of materials thereby requires theoretical and computational tools that span from the atomic-scale detail of first-principles methods to the more coarse-grained description provided by continuum equations with the ultimate aim to systematically couple the scales from the atomistic to the continuum level in [215].

Multi-scale mechanics of nano-composites including an interface to better understand the phenomenological changes across multiple length and time scales are reviewed in [216] and current multi-scale modeling and simulation of nano-structured materials in [217]. Modelling the nano-scale phenomena in condensed matter physics via computer-based numerical simulations focusing on the adhesive and indentation properties of the solid surfaces in nano-contacts, the nucleation and growth of nanophase metallic and semi-conducting atomic and molecular films on supporting substrates, and the nano- and multiscale crack propagation properties of metallic lattices are discussed in [218]. Heterogeneous multiscale methods, applicable on complex fluids, micro-fluidics, solids, interface problems, stochastic problems, and statistically self-similar problems are reviewed in [219] and the emerging role of multiscale modeling in nano- and micro-mechanics of materials in [220].

Hierarchical modeling of amorphous polymers where a broad spectra of length and time scales governing the behavior of these materials is based on the use of connectivity-altering Monte Carlo algorithms for rapid equilibration of atomistic models of long-chain polymer systems, calculation of their conformational, packing and volumetric properties, and assessment of their entanglement structure, and self-consistent field calculations of morphology development in complex systems containing block copolymers, coupled with rubber elasticity theory for the prediction of the stress-strain behavior of these systems are discussed in [221]. Linking various length scales via materials informatics is reviewed in [222]. Recent progress in the simulations of liquid crystals across a range of length and time scales is reviewed in [223] with three material properties of liquid crystals (the archetypal self-assembled materials in Nature) are discussed in detail: elastic constants, rotational viscosity and helical twisting powers. Computer simulations of surfactant solutions are reviewed by Shelley & Shelley discussing the importance of connecting the inherently disparate length scales treating the systems using a single coherent multiscale simulation [224]. In an excellent review dealing with self-assembly from nano-scale to micro-scale colloids, issues like the effect of the shape and composition on assembly and the role of these factors in the self-organization of particles into ordered assemblies, scale dependent effects on assembly related to interparticle forces, vitrification and gelation and also building block design rules from computer simulations using anisotropic interactions are discussed in [225]. Karakasisis and Charitidis give an excellent review of methods in multiscale computational materials science how to connect electronic structure calculations with classical atomistic simulation using molecular dynamics or Monte Carlo methods at the nano/micro scale and further with kinetic Monte Carlo for larger system/time scales and finite elements for very large scales, presenting both hierarchical and hybrid strategies in [226]. A multiscale modelling scheme is reviewed, starting the initial nucleation processes using MD simulation and inputing the data into a cellular automaton (CA)-based model of the microstructure formation at the micro-scale, as well as to the macroscopic heat flow equation in [227]. Thermodynamics of self-assembly of surfactants in solution through simulations is now being expanded to include phenomena in the fluid dynamic regime a smooth link from MD to mesoscopic and macroscopic length and time scales. Recent trends in this area along with new results based on classical approaches are reviewed in [228].

Future Directions

Self-assembly is a process of pivotal importance that should be very attractive to study using modelling and simulations. The intermolecular interactions that drives the process can be directly introduced into common forms of the interaction potentials and force-fields. Many self-assembly processes can indeed also be rationalized and understood both molecularly and macroscopically by means of thermodynamics.

Self-assembly is typically a relatively slow process that proceeds through several metastable states. The extended time-scale and structure evolution that usually extends of several length scales requires a combination of approaches. The rapid expansion of multi scale modeling approaches show much promise but it is still a long way to go before the fine details of self assembly processes with durations of hours up to weeks can be captured accurately. In principle the kinetic Monte Carlo method could cover arbitrary long time scales but it requires detailed knowledge about the system, its states and barriers between them.

Another great challenge in modern computational molecular science is the transition from a descriptive to a predictive level of molecular modeling. Despite several examples of successful predictions of molecular properties *in silico*, current molecular modeling remains largely on a descriptive stage: simulations are used to describe already known experimental phenomena, in order to give a better understanding, for testing the theories describing experiment, or for correct interpretation of the experimental data. Ability of simulations to predict, to discover something new is still rather limited. While some properties are predicted, others are not, successful investigations are published while failures are not, etc., so it is still difficult to rely on simulation results and experimental verification of simulation is therefore always needed. One of the reasons to a limited predictability of molecular models is that the currently used force fields were developed originally at the early stages of molecular modeling in the 80s with rather limited computing resources at hands and later only refined in a few occasions. In the case of meso-scale modeling on a coarse-grained level, even a functional form of the interaction potential is not known and therefore many *ad hoc* models are used. In order to substantially improve reliability and predictability of molecular simulations, a new generation of force fields would be necessary, which have to be derived in a more consecutive manner with more emphasis on the fundamental *ab-initio* approaches. To replace experiment with a computer experiment surely lies far in the future. However, if modelling can point the right direction to an experimentalist, much guesswork, effort and resources can be saved.

Acknowledgments

Both authors are connected to the Berzelii centre EXSELENT on porous materials at Stockholm University (http://www.exselent.su.se/).

Bibliography

1. Ozin GA, Arsenault AC (2005) Nanochemistry, A Chemical Approach to Nanomaterials. RSC Publishing, London
2. Whitesides GM, Grzybowski B (2002) Science 295:2418
3. Lehn J-M (1988) Angew Chem Int Ed 27:89
4. Nuzzo RG, Fusco FA, Allara DL (1987) J Am Chem Soc 109:2358
5. Bentzon MD, van Wonterghem J, Moerup S, Tholen A, Koch CJW (1989) Phil Mag B 60:169
6. Steigerwald ML, Alivisatos AP, Gibson JM, Harris TD, Kortan R, Muller AJ, Thayer AM, Duncan TM, Douglass DC, Brus LE (1988) J Am Chem Soc 110:3046
7. Redl FX, Cho K-S, Murray CB, O'Brien S (2003) Nature 423:968
8. Shevchenko EV, Talapin DV, Kotov NA, O'Brien S, Murray CB (2006) Nature 439:55
9. Decher G (1997) Science 277:1232
10. Whitesides GM, Mathias JP, Seto CT (1991) Science 254:1312
11. Fendler JH (1996) Chem Matter 8:1616
12. Lawrence DS, Jiang T, Levett M (1996) Chem Rev 95:2229
13. Zhang S (2003) Nature Biotech 21:1171
14. Tirrell M (2005) AIChE J 51:2386
15. Zhao X, Zhang S (2006) Chem Soc Rev 35:1105
16. Dubois LH, Nuzzo RG (1992) Ann Rev Phys Chem 43:437
17. Ulman A (1996) Chem Rev 96:1533
18. Laibinis PE, Whitesides GM, Allara DL, Tao YT, Parikh AN, Ruzzo RG (1991) J Am Chem Soc 113:7152
19. Tanev PT, Pinnavaia TJ (1995) Science 267:865
20. Zeng FW, Zimmerman SC (1997) Chem Rev 97:1681
21. Lyubartsev AP, Laaksonen A (2003) In: Karttunen M, Vattulainen I, Lukkarinen A (eds) Novel methods in soft matter simulations. Lecture Notes in Physics, vol 640. Springer, Berlin
22. Laaksonen A, Tu Y (1999) In: Balbuena PB, Seminario JM (eds) Molecular Dynamics: from classical to quantum methods. Elsevier, Amsterdam
23. Roothaan CCJ (1951) Rev Mod Phys 23:69
24. Hehre WJ, Radom L, Schleyer PVR, Pople JA (1986) Ab initio molecular orbital theory. Wiley, New York
25. Leach A (2001) Molecular Modelling: Principles and Applications, 2nd edn. Pearson Education, New Jersey
26. Pople JA, Santry DP, Segal GA (1965) J Chem Phys 43:S129
27. Pople JA, Segal GA (1965) J Chem Phys 43:S136
28. Pople JA, Beveridge DL, Dobosh PA (1967) J Chem Phys 47:2026
29. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP (1985) J Am Chem Soc 107:3902
30. Stewart JJP (1989) J Comput Chem 10:209
31. Stewart JJP (1989) J Comput Chem 10:221
32. Dewar MJS, Thiel W (1977) J Am Chem Soc 99:4899
33. Kolb M, Thiel W (1993) J Comput Chem 14:775
34. Weber W, Thiel W (2000) Theor Chem Acc 103:495
35. Tu Y, Jacobsson SP, Laaksonen A (2003) Mol Phys 101:3009
36. Lee C, Yang W, Parr RG (1988) Phys Rev B 37:785
37. Burke K, Perdew JP, Wang Y (1998) In: Dobson JF, Vignale G, Das MP (eds) Electronic Density Functional Theory: Recent Progress and New Directions. Plenum Press, New York
38. Becke AD (1988) Phys Rev A 38:3098
39. Hohenberg P, Kohn W (1964) Phys Rev B 136:864
40. Kohn W, Sham LJ (1965) Phys Rev A 140:1133
41. Car RR, Parrinello M (1985) Phys Rev Lett 55:2471
42. Foulkes WMC, Haydock R (1989) Phys Rev B 39:12520
43. Harris J (1985) Phys Rev B 31:1770
44. Sankey OF, Niklewski DJ (1989) Phys Rev B 40:3979
45. Tu Y, Laaksonen A (2005) In: Chipot C, Elber R, Laaksonen A, Leimkuhler B, Mark A, Schlick T, Schuette C, Skeel R (eds) Advances in Algorithms for Macromolecular Simulation. Lecture Notes in Computational Science & Engineering, vol 49. Springer, Heidelberg
46. Tu Y, Jacobsson SP, Laaksonen A (2006) Phys Rev B 74:205104

47. Madden PA, Heaton R, Aguado A, Jahn S (2006) J Mol Struct – Theochem 771:9
48. Hafner J (2007) Comp Phys Commun 177:6
49. Muller S (2006) Surf Interf Anal 38:1158
50. Juhas P, Cherba DM, Duxbury PM, Punch WF, Billinge SJL (2006) Nature 440:655
51. Le Page Y, Rodgers JR (2005) Acta Appl Cryst 38:697
52. Bockstedte M, Kley A, Neugebauer J, Scheffler M (1997) Comp Phys Commun 107:187
53. Gonze X, Beuken JM, Caracas R, Detraux F, Fuchs M, Rignanese GM, Sindic L, Verstraete M, Zerah G, Jollet F, Torrent M, Roy A, Mikami M, Ghosez P, Paty JY, Allan DC (2002) Comp Mater Sci 25:478
54. Colombo L (2005) Riv Nuovo Cim 28:1
55. Bowler DR, Choudhury R, Gillan MJ, Miyazaki T (2006) Phys Stat Solidi B 243:989
56. Kawazoe Y (2003) Bull Mater Sci 26:13
57. Allen MP, Tildesley D (1987) Computer Simulation of Liquids. Clarendon, Oxford
58. Rapaport DC (2004) The Art of Molecular Dynamics Simulation, 2nd edn. Cambridge University Press, Cambridge
59. Frenkel D, Smit B (2002) Understanding Molecular Simulation: from Algorithms to Applications, 2nd edn. Computational Science Series, vol 1. Academic Press, London
60. Schlick T (2002) Molecular Modeling: An Interdisciplinary Guide. Springer, New York
61. Hinchliffe A (2003) Molecular Modelling for Beginners. Wiley, New York
62. Field MF (2007) A Practical Introduction to the Simulation of Molecular Systems. Cambridge University Press, Cambridge
63. Seminario JM, Balbuena P (eds) (1999) Molecular Dynamics (Theoretical and Computational Chemistry). Elsevier Science, Amsterdam
64. Haile JM (1992) Molecular Dynamics Simulation: Elementary Methods. Wiley, New York
65. Leimkuhler B, Reich S (2004) Simulating Hamiltonian Dynamics. Cambridge University Press, Cambridge
66. Alder BJ, Wainwright TE (1957) J Chem Phys 27:1208
67. Alder BJ, Wainwright TE (1959) J Chem Phys 31:459
68. Rahman A (1964) Phys Rev 136:405
69. Verlet L (1967) Phys Rev 159:98
70. Stillinger FH, Rahman A (1974) J Chem Phys 60:1545
71. Hedman F, Laaksonen A (1999) In: Balbuena PB, Seminario JM (eds) Molecular Dynamics: from classical to quantum methods. Theoretical Chemistry Series, vol 7. Elsevier Science BV, Amsterdam
72. Hedmann F (2006) http://www.diva-portal.org/diva/getDocument?urn_nbn_se_su_diva-1008-4__fulltext.pdf
73. Sprik M, Demarchee E, Michel B, Rothlisberger U, Klein ML, Wolf H, Ringsdorf H (1994) Langmuir 10:4116
74. Yamamoto H, Watanabe T, Nishiyama K, Tatsumura K, Ohdomari I (2006) J Phys IV 132:189
75. Mamatkulov SI, Khabibullaev PK, Netz TR (2004) Langmuir 20:4756
76. Yu KQ, Li ZS, Sun JZ (2002) Langmuir 18:1419
77. Hackett E, Manias E, Giannelis EP (1998) J Chem Phys 108:7410
78. Jang SS, Jang YH, Kim Y-H, Goddard III WA, Flood AH, Laursen BW, Tseng H-R, Stoddart JF, Jeppesen JO, Choi JW, Steuerman DW, Delonno E, Heath JR (2005) J Am Chem Soc 127:1563
79. Striolo A, NcCabe C, Cummings PT (2006) J Chem Phys 125:104904
80. Luo M, Mazyar OA, Zhu Q, Vaughn MW, Hase WL, Dai LL (2006) Langmuir 22:6385
81. Wu J (2006) AIChE J 52:1169
82. Jallipalli A, Balakrishnan G, Huang SH, Khoshakhlagh A, Dawson LR, Huffaker DL (2007) J Cryst Growth 303:449
83. van Workum K, Douglas JF (2006) Phys Rev E 73:031502
84. Ariga K, Hill JP, Ji Q (2007) Phys Chem Chem Phys 9:2319
85. Cooper N (1987) Stanislaw Ulam 1909–1984. Los Alamos Sci, vol 15. Special Issue. See: http://www.fas.org/sgp/othergov/doe/lanl/pubs/number15.htm
86. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AN, Teller E (1953) J Chem Phys 21:1087
87. Jayaram B (1996) J Math Chem 20:395
88. Ni R, Cao D, Wang WC (2007) J Phys Chem C 111:11802
89. Pool R, Schapotschnikow P, Vlugt TJH (2007) J Phys Chem 111:10201
90. Sinyagin AY, Belov A, Tang Z, Kotov NA (2006) J Phys Chem B 110:7500
91. Reimer U, Zehl T, Wahab M, Schiller P, Moegel H-J (2006) Coll Surf A Phys Chem Eng Asp 290:25
92. Reimer U, Wahab M, Schiller P, Moegel HJ (2006) Langmuir 21:1640
93. Patti A, Mackie AD, Siperstein FR (2007) Langmuir 23:6771
94. He XH, Song M, Liang HJ, Pan CY (2001) J Chem Phys 114:10510
95. Wijmans C, Linse P (1995) Langmuir 11:3748
96. Wijmans C, Linse P (1996) J Phys Chem 100:12583
97. Wijmans C, Linse P (1997) J Chem Phys 106:328
98. Linse P (2000) In: Alexandridis P, Lindman B (eds) Amphiphilic Block Copolymers: Self-Assembly and Application. Elsevier Science, Amsterdam
99. Abraham FF, White GM (1970) J Appl Phys 41:1841
100. Gilmer GH, Bennema P (1972) J Appl Phys 43:1347
101. Bortz AB, Kalos MH, Lebowitz JL (1975) J Comput Phys 17:10
102. Gillespie DT (1977) J Phys Chem 81:2340
103. Chatterjee A, Vlachos DG (2007) J Computer-Aided Matter Des 14:253
104. Metiu H, Lu Y-T, Zhang Z (1992) Science 255:1088
105. Much F, Ahr M, Biehl M, Kinzel W (2001) Europhys Lett 56:791
106. Larsson MI (2004) Surf Sci 551:69
107. Kozak JJ, Nicolis C, Nicolis G (2007) J Chem Phys 126:154701
108. Zhu R, Pan E, Chung PW (2007) Phys Rev B 75:205339
109. JS Rowlinson Notes (2003) Rec R Soc Lond 57:143
110. Wheeler AA, Boettinger WJ, McFadden GB (1992) Phys Rev A 45:7424
111. Anderson DM, McFadden GB, Wheeler AA (1998) Ann Rev Fluid Mech 30:139
112. Lu W, Suo Z (2001) J Mech Phys Solids 49:1937
113. Chen LQ (2002) Annu, Rev Mater Res 32:113
114. Boettinger WJ, Warren JA, Beckermann C, Karma A (2002) Ann Rev Mater Res 32:163
115. Thornton K, Ågren J, Voorhees PW (2003) Acta Mater 51:5675
116. Beckermann C, Diepers H-J, Steinbach I, Karma A, Tong X (1999) J Comput Phys 154:468
117. Boettinger WJ, Coriell SR, Greer AL, Karma A, Kurz W, Rappaz M, Trivedi R (2000) Acta Mater 48:43
118. Cahn JW, Hillard JE (1957) J Chem Phys 28:258
119. Cahn JW, Hillard JE (1959) J Chem Phys 31:688
120. Cahn JW (1961) Acta Metall 9:795

121. Allen SM, Cahn JW (1979) Acta Metall Mater 27:1085
122. Chen LQ, Shen J (1998) Comput Phys Commun 108:147
123. Elder KR, Grant M (2004) Phys Rev E 70:051605
124. Lu W, Salac D (2005) Phys Rev Lett 94:146103
125. Wang YU (2007) Acta Mater 55:3835
126. Inoue T (2004) J Phys IV 120:3
127. Inoue T (2004) Mater Sci Res Intl 10:1
128. Hoyt JJ, Sadigh B, Asta M, Foiles SM (1999) Acta Mater 47:3181
129. Gránásy L, Börsönyi T, Pusztai T (2002) Phys Rev Lett 88: 206105
130. Gránásy L, Börsönyi T, Pusztai T (2002) J Cryst Growth 237–239:1813
131. Gránásy L, Pusztai T, James PF (2002) J Chem Phys 117:6157
132. Gránásy L, Pusztai T, Warren JA, Douglas JF, Börsönyi T, Ferrero V (2003) Nat Mater 2:92
133. Kvamme B, Graue A, Aspenes E, Kuznetsova T, Gránásy L, Tóth G, Pusztai T, Tegze G (2004) Phys Chem Chem Phys 6:2327
134. Gránásy L, Pusztai T, Börsönyi T, Warren JA, Douglas JF (2004) Nat Mater 3:645
135. Svandal A, Kvamme B, Gránásy L, Pusztai T (2005) J Phase Equil Diff 26:534
136. Pusztai T, Bortel G, Gránásy L (2005) Mater Sci Eng A 413–414:412
137. Svandal A, Kvamme B, Gránásy L, Pusztai T, Buanes T, Hove J (2006) J Cryst Growth 486:1813
138. Tegze G, Pusztai T, Tóth G, Gránásy L, Svandal A, Buanes T, Kuznetsova T, Kvamme B (2006) J Chem Phys 124:234710
139. Gránásy L, Pusztai T, Saylor D, Warren JA (2007) Phys Rev Lett 98:035703
140. Davidchack RL, Laird BB (1998) J Chem Phys 108:9452
141. Davidchack RL, Laird BB (2000) Phys Rev Lett 85:4751
142. Plapp M, Karma A (2000) J Comput Phys 165:592
143. Raabe D (2002) Adv Mater Progr Rept 14:639
144. Penrose MD (2001) Commun Math Phys 218:153
145. Bartelt MC, Privman V (1991) Intern J Mod Phys B5:2883
146. Evans JW (1993) Rev Modern Phys 65:1281
147. Privman V (2000) Colloids Surf A 165:231
148. Talbot J, Tarjus G, van Tassel PR, Viot P (2000) Colloids Surf A 165:287
149. Senger B, Voegel J-C, Schaaf P (2000) Colloids Surf A 165:255
150. Rényi A (1963) Trans Math Stat Prob 4:205
151. D'Orsogna MR, Chou T, Antal T (2007) J Phys A Math Theor 40:5575
152. Ranjith P, Marko JF (2006) Phys Rev E 74:041602
153. Adamczyk Z, Weroński P, Barbasz J, Kolańsiska M (2007) Appl Surf Sci 253:5776
154. Gray JJ, Klein DH, Korgel BA, Bonnecaze RT (2001) Langmuir 17:2317
155. Breit M, Gao M, von Plessen G, Lemmer U, Feldmann J, Cundiff ST (2002) J Chem Phys 117:3956
156. Yuan Y, Oberholzer MR, Lenhoff AM (2000) Colloids Surf A165:125
157. Erban R, Chapman SJ (2007) Phys Rev E 75:041116
158. Cadilhe A, Aráujo N, Privman V (2007) J Phys Condens Matter 19:065124
159. Bohidar HB, Mohanty B (2004) Phys Rev E 69:021902
160. Gray JJ, Bonnecaze RT (2001) J Chem Phys 114:1366
161. Wolfram S (2002) A new kind of science. Wolfram Media, Champaign
162. Petri CA (1962) Kommunikation mit Automaten. Ph D Thesis, University of Bonn
163. Turing AM (1952) Phil Trans R Soc Lond B237:37
164. Wishart DS, Yang R, Arndt D, Tang P, Cruz J (2005) In Silico Biol 5:139
165. Succi S (2001) The Lattice Boltzmann Equation for Fluid Dynamics and Beyond. Oxford University Press, Oxford
166. Tougaw PD, Lenta CS (1996) J Appl Phys 80:4722
167. Amlani I, Orlov AO, Snider GL, Lent CS, Bernstein GH (1998) Appl Phys Lett 72:2179
168. Orlov AO, Amlani I, Bernstein GH, Lent CS, Snider GL (1997) Science 277:928
169. Snider GL, Orlov AO, Amlani I, Zuo X, Bernstein GH, Lent CS, Merz JL, Porod W (1999) J Appl Phys 85:4283
170. Cole T, Lusth JC (2001) Progr Quant Electr 25:165
171. Lent CS, Isaksen B, Lieberman M (2003) J Am Chem Soc 125:1056
172. Israeli N, Goldenfeld N (2006) Phys Rev E 73:026203
173. Mayer B, Köhler G, Rasmussen S (1997) Phys Rev E 55:4489
174. Nilsson M, Rasmussen S (2003) Discret Math Theor Comput Sci AB(DMCS):31
175. Kier LB, Cheng C-K, Testa B, Carrupt P-A (1997) J Pharmac Sci 86:774
176. Sreekumar P, Jayaraman VK, Kulkarni BD (1998) Ind Eng Chem Res 37:2188
177. Demontis P, Pazzona FG, Suffritti GB (2007) J Chem Phys 126:194709
178. Gumbsch P (1995) J Mater Res 10:2897
179. Lyubartsev AP, Laaksonen A (1995) Phys Rev E 52:3730
180. Muller-Plathe F (2002) Chem Phys Chem 3:754
181. Izvekov S, Parrinello M, Burnham CJ, Voth GA (2004) J Chem Phys 120:10896
182. Du MH, Cheng HP (2003) Int J Quant Chem 93:1
183. Mallik A, Taylor DCE, Runge K, Dufty JW (2004) Int J Quant Chem 100:1019
184. Bartlett RJ, McClellan J, Greer JC, Monaghan S (2006) J Comp Aided Mater Des 13:89
185. Zhu WM, Runge K, Trickey SB (2006) J Comp Aided Mater Des 13:75
186. Faller R (2004) Polymer 45:3869
187. Buehler MJ, Hartmaier A, Gao H (2004) Model Sim Mater Sci 12:S391
188. Doi M (2001) J Comp Appl Math 149:13
189. Kinjo T, Hyodo S (2007) Mol Simul 33:417
190. Hyodo S (2007) Mol Sim 33:279
191. Haslam AJ, Moldovan D, Phillpot SR, Wolf D, Gleiter H (2002) Comp Mater Sci 23:15
192. Nakano A (1997) Comp Phys Commun 105:139
193. Shen LM, Chen Z (2005) Int J Solids Struct 42:5036
194. Curtin WA, Miller RE (2003) Model Simul Mater Sci Eng 11:R33
195. Rafii-Tabar H, Hua L, Cross M (1998) J Phys Cond Matter 10:2375
196. Smit B, Krishna R (2003) Chem Eng Sci 58:557
197. Warrier M, Schneider R, Salonen E, Nordlund K (2004) Contr Plasma Phys 44:307
198. Shibuta Y, Elliott JA (2006) Chem Phys Lett 427:365
199. Sakaguchi N, Watanabe S, Takahashi H, Faulkner RG (2004) J Nucl Mater 329:1166
200. Vasenkov AV, Fedoseyev AI, Kolobov VI, Choi HS, Hong KH, Kim K, Kim J, Lee HS, Shin JK (2006) J Comp Theor Nanosci 3:453

201. Zhang J, Adams JB (2004) Comp Mater Sci 31:317
202. La Magna A, Coffa S, Libertino S, Brambilla L, Alippi P, Colombo L (2001) Nucl Instr Meth Phys Res B 178:154
203. Haran M, Goose JE, Clote NP, Clancy P (2007) Langmuir 23:4897
204. Netz RR (2004) J Phys Cond Mat 16:S2353
205. Ma J, Lu H, Wang B, Roy S, Hornung R, Wissink A, Komanduri R (2005) Comp Model Eng Sci 8:135
206. Izvekov S, Violi A (2006) J Chem Theory Comput 2:504
207. Izvekov S, Violi A, Voth GA (2005) J Phys Chem B Lett 109:17019
208. Akkermans RLC, Warren PB (2004) Phil Trans Royal Soc A 362:1783
209. Liu WK, Karpov EG, Zhang S, Park HS (2004) Comp Meth Appl Mech Eng 193:1529
210. Starrost F, Carter EA (2002) Surf Sci 500:323
211. Lium WK, Park HS, Qian D, Karpov EG, Kadowaki H, Wagner GJ (2006) Comp Meth Appl Mech Eng 195:1407
212. Csanyi G, Albaret T, Moras G, Payne MC, De Vita A (2005) J Phys Cond Matter 17:R691
213. Rudd E, Broughton JQ (2000) Physica, Status Solidi B – Basic Res 217:251
214. Nieminen RM (2002) J Phys Cond Matter 14:2859
215. Vvedensky DD (2004) J Phys Condens Matter 16:R1537
216. Buryachenko VA, Roy A, Lafdi K, Anderson KL, Chellapilla S (2005) Compos Sci Tech 65:2435
217. Gates TS, Odegard GM, Frankland SJV, Glancy TC (2005) Compos Sci Tech 65:2416
218. Rafii-Tabar H (2000) Phys Rept Rev Sect Phys Lett 325:240
219. Weinan E, Engquist B, Li X, Ren WQ, vanden-Eijnden E (2007) Commun Comp Phys 2:367
220. Ghoniem NM, Cho K (2002) Comp Model Eng Sci 3:147
221. Theodorou N (2005) Comp Phys Commun 169:82
222. Liu ZK, Chen LQ, Rajan K (2006) JOM 58:42
223. Wilson M (2005) Intl Rev Phys Chem 24:421
224. Shelley JC, Shelley MY (2000) Curr Opin Coll Interf Sci 5:101
225. Glotzer SC, Solomon MJ, Kotov NA (2004) AIChE J 50:2978
226. Karakasidis TE, Charitidis CA (2007) Mater Sci Eng C 27:1082
227. Rafii-Tabar H, Chirazi A (2002) Phys Rep 365:145
228. Rajagopalan R (2001) Curr Opin Coll Interf Sci 6:357

Self-Organization and the City

Juval Portugali
ESLab (Environmental Simulation Lab), Department
of Geography and the Human Environment,
Tel Aviv University, Tel Aviv, Israel

Article Outline

Glossary

Self organization A property of open and complex systems that achieve their order spontaneously, that is, by means of "self-organization".

City A form of settlement that first emerged in the Near East (the core being Mesopotamia) some 5500 years ago. Since its first appearance in Mesopotamia it has diffused in space and time. With colonialism the Western–European form of city has diffused to the entire world, suppressing on the way other forms of cities (e. g. in South America, East Asia, etc.). In the last few decades the city has become the most dominant form of settlement: for the first time in human history more than half of world population lives in cities.

Urbanism The term refers to the totality of life in cities: the interrelations between the social structure, culture, economy, politics, architecture, physical morphology, … associated with life in cities. The first appearance of cities is thus termed *the urban revolution*. Most students of urbanism would agree that 21st century society is undergoing a major urban transformation; some describe it as a new urban revolution.

Planning Planning is, on the one hand, a basic cognitive capability of humans while on the other, a profession and research domain termed interchangeably *city planning*, *urban and regional planning* or *environmental planning*. City planners work and act in the context of planning law and administration the aim of which is to regulate and control life in cities.

SIRN (synergetic inter-representation network) An approach to cognitive mapping and urban dynamics suggesting that cities emerge, maintain their order and change again as a consequence of an on-going interaction between cognitive maps that are constructed in the mind/brain of humans as internal representations and the city as a collective external representation. This on-going interaction gives rise to a network some of whose elements are in the mind/brain, while others in the world.

Information compression, inflation and adaptation A view suggesting that Shannon's notion of information is a property of closed systems and that in complex, self-organizing systems one has to take into consideration the role of *semantic information*. Due to semantic information, the process of self-organization often entails *information compression*; in some cases it entails *information inflation*. The suggestion is that information compression and inflation are two facets of the process of *information adaptation*.

Definition of the Subject

Self-organization is a central property of open and complex systems. While the concept had already appeared in the 1940s, its modern use was pioneered in the 1960s by people such as Haken [50,51] with his theory of *synergetics*, Prigogine with his notion of *dissipative structures* [87,109,110] and others (see review in Chap. 3.1 [95]). Such systems are typically in "a far from equilibrium condition" and exhibit phenomena of chaos, fractal structure and the like. For a long time, the term "self-organization" was used also as an umbrella name for these theories; nowadays it is common to refer to these theories as *complexity theories*.

The notions of self-organization and complexity originated in the sciences, specifically in physics, as a property of natural systems. However, as we shall see below, from the start they were associated with the city – at the beginning the city was used as a metaphor to convey the notion of "self-organization" [87] and at a later stage it was studied as a genuine self-organization system in its own sake [5].

Most theories and methodologies of complexity developed in the last three decades have been applied also to the study of cities with the result that we now have a rich body of research on *fractal cities* [20], *self-organization and the city* [95], *cities and complexity* [16], cellular automata and agent base urban simulation models [22], studies on cities from the perspective of Bak's self-organized criticality [20], studies on cities as networks [16] and much more. This growing body while enriching our understanding of cities and providing sophisticated tools to city planning, also exposes problems that will become the challenges for the next generation of studies on complexity, self-organization and the city.

Introduction

The title "Self-Organization and the City" enfolds *two notions* – 'Self-organization' and 'City', *a thesis* suggesting that cities are complex self-organizing systems, and an *inconsistency* – the view of cities as complex systems that achieve their order spontaneously contradicts the traditional view of cities as symbols of organized order and planning. From the title thus follow four questions: what is self-organization? What is a city? In what sense are cities self-organizing systems? What is the meaning of planning in a self-organizing system? In an *Encyclopedia of Complexity and System Sciences* such as this, there is no need to introduce the term self-organization beyond what has been said about it above; we are thus left with three introductory tasks, namely, to clarify the notion "City", elaborate the thesis that cities are self-organizing systems, and, solve the contradiction between self-organization and planning in the realm of cities.

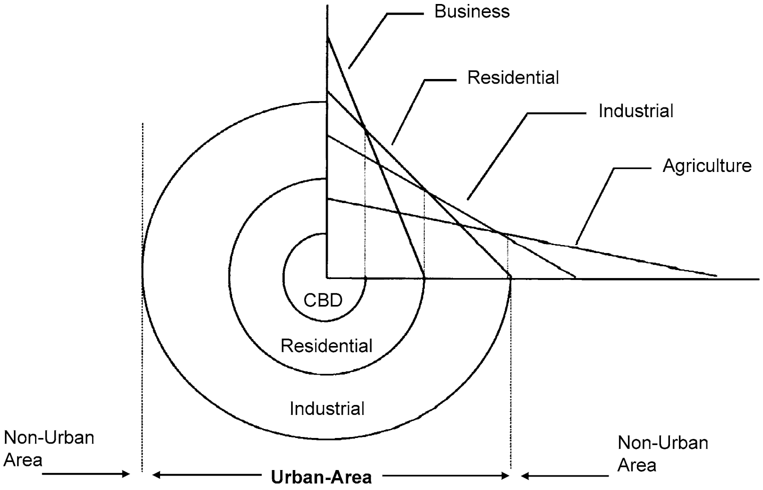
What is a City?

There are two ways to answer this question: first, by looking at explicit attempts to define a city. Second, by exposing the way the different theories of cities explicitly describe, or implicitly perceive, a city.

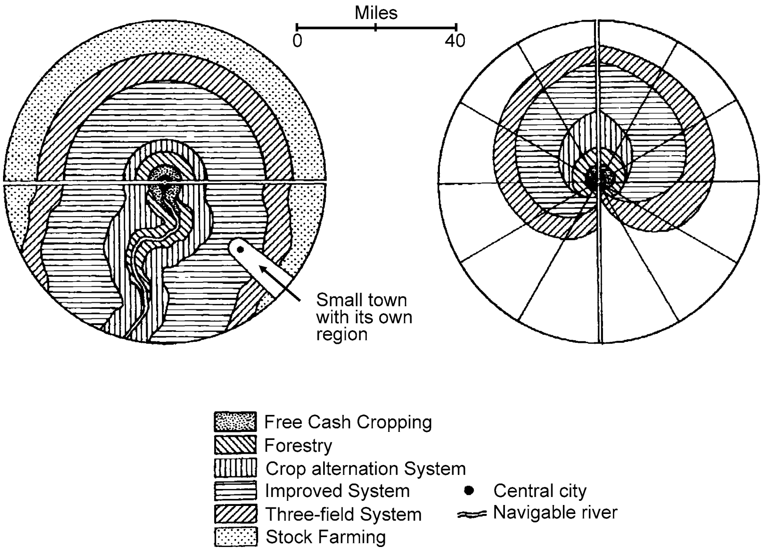
Explicit Attempts to Define a City The history of the many attempts to define "a city" is rather confusing: Whenever a definition was proposed, it was always possible to falsify it (in Popper's [90] sense) by putting forward cities that do not comply with the definition. The main reason for the failure to define cities is that the various attempts to do so were always made with reference to what in cognitive science [115] is called *classical categories*. That is, groups composed of entities sharing some necessary and sufficient conditions that define them as a category and distinguish them from other categories. Students of urbanism have implicitly treated cities as classical categories, and yet, cities are not classical categories – they form a category due to what Wittgenstein [136] has termed *family resemblance*. As a consequence, attempts to define them in terms of a classical category ended up with a failure [95].

A family resemblance category becomes a category not when its elements share some common denominators, but when they form a *network* of partial links and similarities. Further research and experiments have found that many family resemblance categories have a *core-periphery* structure, in the sense that some instances of the category are more *prototypical* of the category than others and they thus form its center while the rest of the instances form the category's periphery [74,78,115].

The city is a good example of a family resemblance category with a core periphery structure. On the one hand, there are no common elements between the "first" cities of some 5,500 years ago and the cities of today except for the name. On the other hand, the first cities had space-time links and similarities with subsequent cities, which in turn had common elements with subsequent cities, and so on until the global cities of today. The result of this process is that cities form a huge space-time family resemblance network extending in time and space from the ancient cities of 5,500 years ago to the cities of today. In this network, one can identify space-time moments during which certain cities became more characteristic or proto-



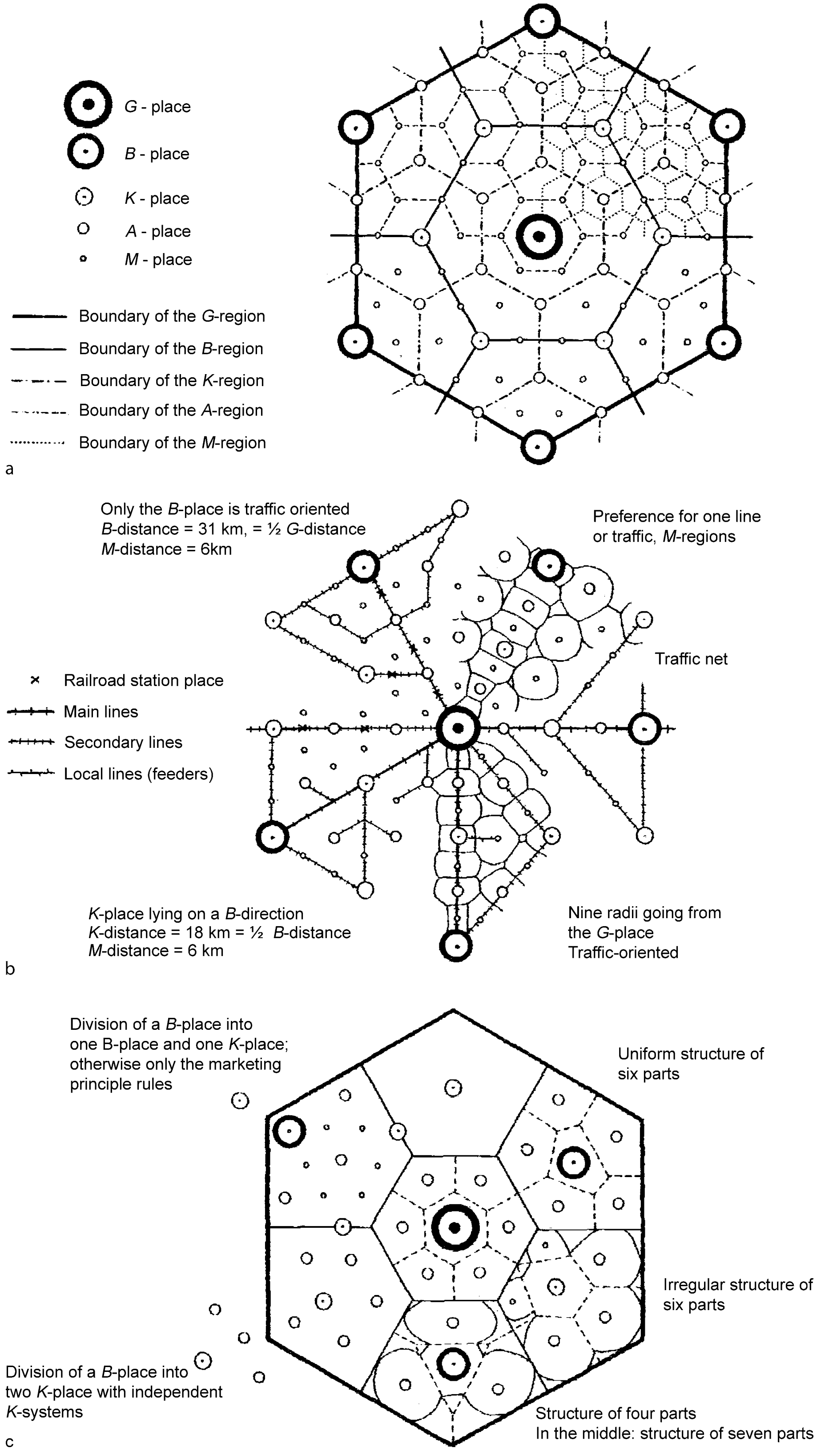
Self-Organization and the City, Figure 1
Thünen's type land-use system as transformed into an urban land-use system by location theory. Businesses are prepared to pay high rent at the center of the city, but are reluctant to "live" far from it. Their spatial demand curves (or rbc – rent bid curves) are thus the highest and steepest. Industrialists, in this exposition, are exactly the opposite and residents are in between: they cannot afford to pay the high prices at the center, but are prepared to live far from it, and so on. Each land use thus occupies a ring were it can pay (bid for) the highest rent. Note that the principle of marginal utility which is implicit in Thünen's landscape, here appears explicitly as the central economic principle



Self-Organization and the City, Figure 2
The diagrams of Thünen's *Isolated State*: *Left, upper part of the diagram* "This shows the Isolated State in the shape it must take from the assumptions made in Section One...". *Left, lower part* "Here we see the Isolated State crossed by a navigable river. Here the ring of crop alternation become very much larger, stretching along the river ... The effect of constructing a highway is similar, ... " (Par. 385). *Right* "The diagram illustrates the effect of the Town grain price on the extension of the cultivated plain" (Par. 386). Source: [129]

typical of the category than others. Such cities have temporarily captured the center of the category city, pushing to the periphery the rest of the instances, only to be replaced in subsequent space-time moments by other prototypical cities, other centers and other peripheries. How does this huge network evolve in time and space? The answer is: "by means of self-organization" [95].

Images of Cities This section discusses images of cities that are implicit or explicit in several of the urban theories.


G - place
B - place
K - place
A - place
M - place
Boundary of the G-region
Boundary of the B-region
Boundary of the K-region
Boundary of the A-region
Boundary of the M-region
a
Only the B-place is traffic oriented
B-distance = 31 km, = ½ G-distance
M-distance = 6km
Preference for one line
or traffic, M-regions
Traffic net
Railroad station place
Main lines
Secondary lines
Local lines (feeders)
K-place lying on a B-direction
K-distance = 18 km = ½ B-distance
M-distance = 6 km
Nine radii going from
the G-place
Traffic-oriented
b
Division of a B-place into
one B-place and one K-place;
otherwise only the marketing
principle rules
Uniform structure of
six parts
Irregular structure of
six parts
Division of a B-place into
two K-place with independent
K-systems
Structure of four parts
In the middle: structure of seven parts
c

◂ **Self-Organization and the City, Figure 3**
Christaller's systems of central places according to the three locational principles. a The marketing regions in a system of central places. b A system of central places developed according to the traffic principle. c A system of central places developed according to the separation principle. (Source: Figs. 2, 4, 6 in [36])

Only theories that facilitate subsequent discussion will be discussed.

The Economic City The city is portrayed as a center-periphery structure with several rings that come into being by an economic competition between land uses that differ in their spatial demand function (Fig. 1). The land uses with the highest and steepest *spatial demand curves* or *rent bid curves* [8,69] capture the central ring – the most accessible area of the city – thus forming the CBD (central business district); the rest of the land uses occupy the peripheral rings. The origin of this city image is von Thünen's [129] *Isolated State* (Fig. 2): Originally formulated as a theory of agricultural land uses it became the founding theory of all location theories including urban land use theories.

The City as a Central Place The city is perceived as a central place that mediates between the city's complementary region and other cities that form a hierarchical network of central places (Fig. 3). The origin of this view is Christaller's [36] *central place theory* that perceived the city as a central place for tertiary activities (a market place, transportation node and administrative center). Losch's [81] central place theory was more ambitious and complicated and portrayed the city as a central place for all production, consumption, transportation and political activities (Fig. 4).

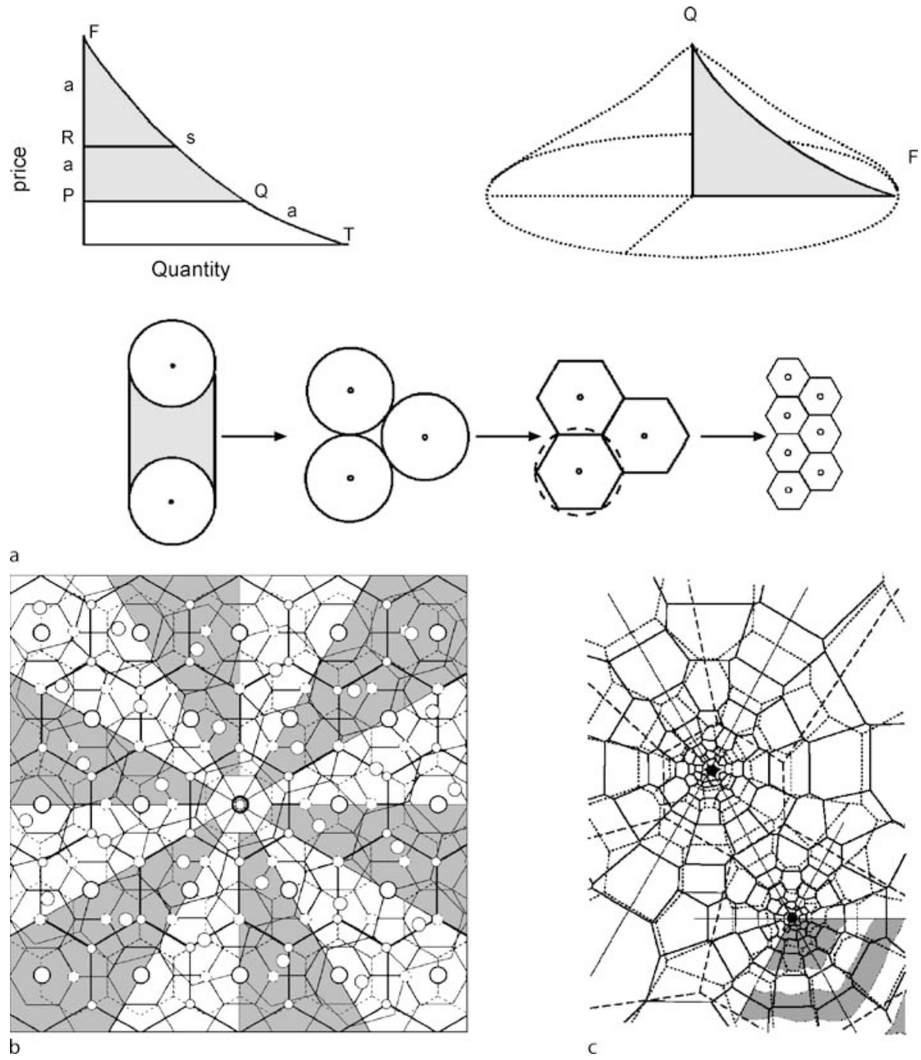
The City as a Node in a System of Cities The city here "looses" its autonomy in the sense that it is perceived as a node in a system of cities – no attention is paid to the city's role or function; the focus of interest is on the system as a whole. This view is due to Auerbach [10] who already at the turn of the 20th century showed that the rank-size distribution of cities obeys the power law. In a famous work from 1949 Zipf showed that this rank-size distribution typifies not only cities (Fig. 5), but a whole range of phenomena [138]. Zipf's work provided a source of inspiration to a long list of subsequent studies on systems of cities [111].

The Ecological City As in the economic city, here too, the city is portrayed as a center-periphery ring structure. However, here the city's structure emerges out of a competition between cultural and socio-economic groups in a way similar to competition between species in natural ecology. The various studies in this domain are termed *urban ecology*; their source of inspiration is the Chicago school of social ecology. Several urban landscapes were suggested the most dominant one is Burgess' [30,31,32] ring structure (Fig. 6).

The City as a Representation of Society The city is here perceived as a spatial representation of society as a whole. This image of the city emerged in the early 1970s as a consequence of a paradigm shift the study of cities underwent – from liberal social and economic theories to more radical ones with Marxism being the most dominant view. Two Marxist interpretations of the city can give the flavor of this approach. The first is Castells' [33] view according to which the city is a spatial representation of the structure of society as perceived by structuralist-Marxist theory (Fig. 7). The second is Harvey's view according to which the city's landscape emerges, as a logical consequence, out of internal contradictions inherent in the capitalist *mode of production* that according to this view, dominates world society of the 20th and 21st centuries. Namely, between forces of spatial agglomeration and processes of spatial dispersion. As illustrated in Fig. 8, this tension can be resolved only by *the urbanization of capital* [47,63].

The City as a Socio-cultural Force The city is here perceived as a force that is shaping the life of the people living in it. In urban societies it implies that the city is in fact shaping society. This view is due to the study of Wirth's [135] *Urbanism as a way of life* and also of Park's study *The City* [88]. In 1970 Lefebvre has published a monograph *La Révolution Urbaine* suggesting from a Marxist point of view that society is reaching a stage of being completely urban so that urbanism is replacing industrialism as the major force of society [79].

The Postmodern City The city of 21st century is described as the postmodern city [95]: Untamed, shrew, capricious, ever-changing; actually it is not a city but a text written by millions of unknown writers, unaware that they are writers, read by millions of readers, each reading his or her own personal and subjective story in this ever-changing chaotic text, thus changing and recreating and further complicating it. Today's urbanism is a big theater at the center of whose stage we see a kaleidoscope of shapes, forms, high-tech science-fiction structures, cultures and sub-cultures, Italians, Chinese, Japanese, Jews, Indians, Gays, Lesbians. Yapese; nothing is stable, nothing



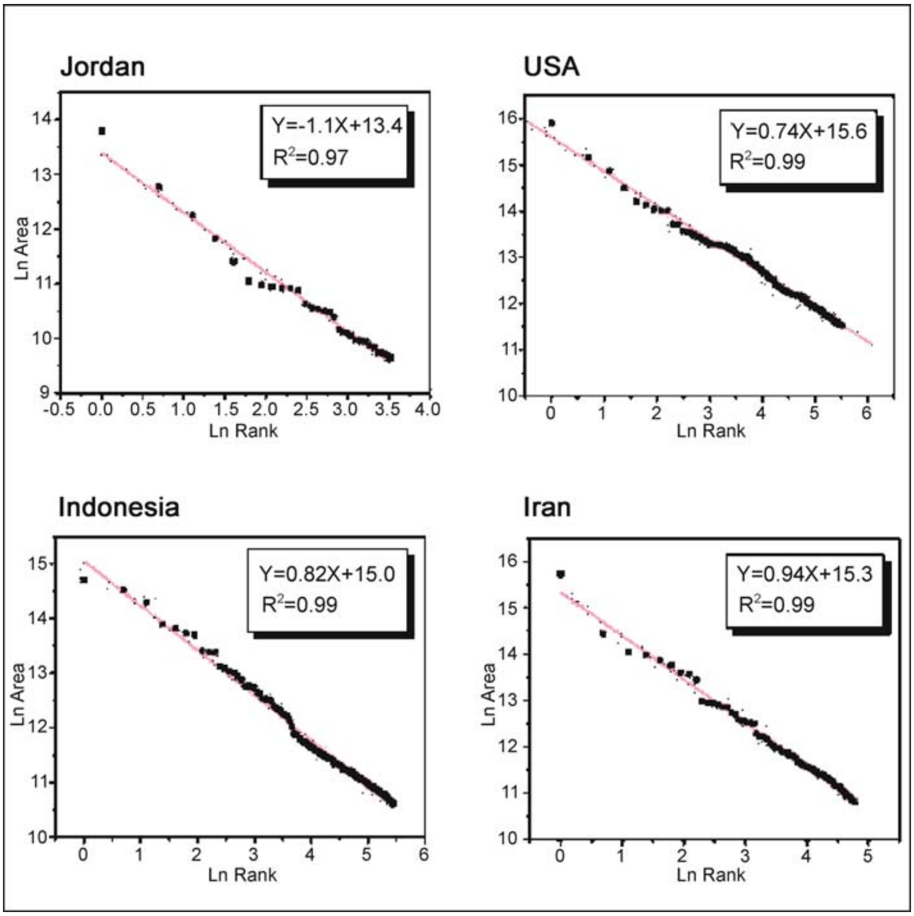
Self-Organization and the City, Figure 4
a The derivation of Lösch's system of central places. *Top* The derivation of a *spatial demand cone* with its market area (*right*) out of an "ordinary" demand curve (*left*). *Bottom* Development of market areas from the large circle to the final small hexagon. Source: Figs. 20–23 in [81]. b Lösch's derived system of central places with their market areas, divided into "city-poor", "city-rich" sectors. Source: Fig. 28 in [81]. c A Lösch system of central places modified by Isard [69] so as to be consistent with the resulting population distribution

is true nor matter for more than a second, not the Marxist urban categories, nor any other grand theory or truth; all must go, must move, clear the way to the new next whatever it is.

The Self-organizing City Strangely enough, an image of the city similar to the postmodern one is emerging out of complexity studies of cities. This is a seemingly similarity, however; a closer look reveals, first, that theories of complexity made a direct and explicit link to the views of cities as central places, to the studies on systems of cities and to the ecological views on the city. Second, that the notion of self-organizing city has several important resemblances with modern social theory oriented urban studies that perceive the city as the representation of society.

In What Sense Are Cities Self-organized Systems?

Self-organization is a property of systems that are open and complex. No one plan such systems, no one fully controls them and yet they have order, rules, organization and all these emerge spontaneously by means of self orga-



Self-Organization and the City, Figure 5

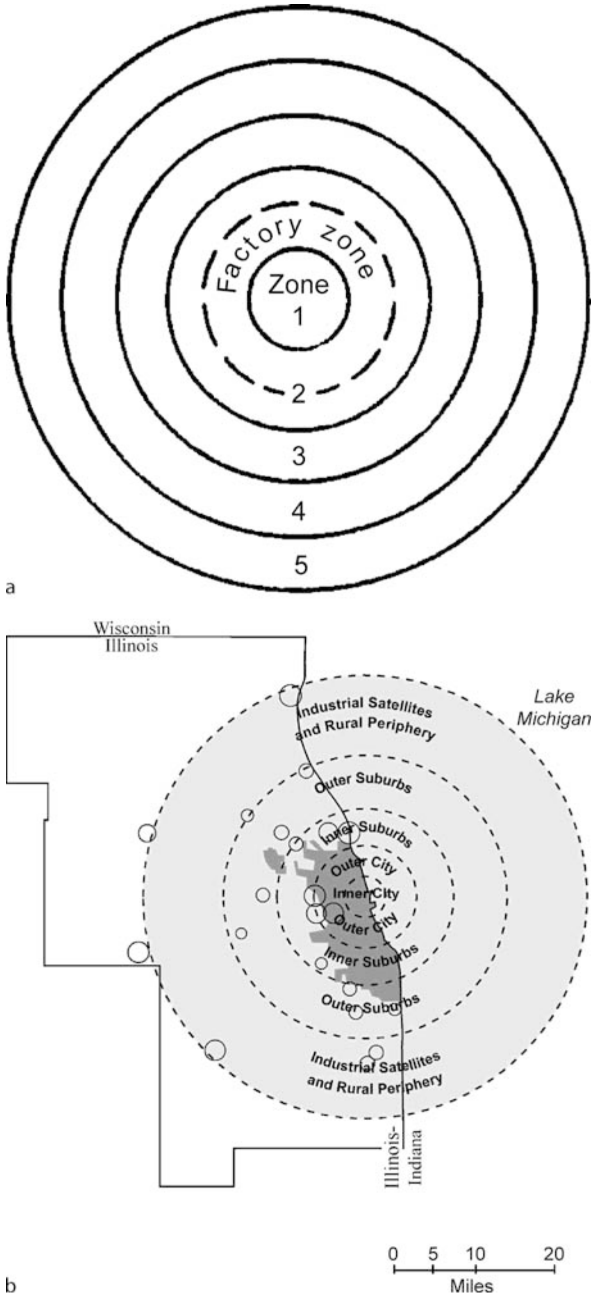
Rank size distribution of cities with more than 100,000 inhabitants in four different countries. Source: Fig. 2.7 in [27]

nization. A nice example of self-organization is provided by human languages (Chinese, Hebrew, English …). Each such language is an open system; each is a complex system; each is a system that emerged out of synergetic interaction between a huge number of people (the "parts" of such systems); no one has ever fully controlled languages; no one has fully planned a language; and yet each of the human languages has order, rules, organization and all these emerged spontaneously by means of self organization.

Similarly to human languages, each city is an artifact; each is an open system; each is a complex system; each city is a system that emerged out of interaction between a huge number of people; no one fully controls it; and yet it has order, rules, organization and all these spontaneously by means of self organization.

But cities are not languages. For one thing, their products are stand-alone objects such as buildings, roads, bridges, etc. that exist and survive independently of their producers. The products of languages are humans' voices and gestures that have no existence independent of their producers. Cities, in this respect, are akin to writing and texts – the external, stand-alone, representations of languages. The appearance of cities, some 5,500 years ago, hand in hand with writing, is, to my mind, not accidental.

A second difference concerns planning: There are no language planners and the attempt to "plan" the international language of Esperanto ended in failure. But there are many city planners – much more than appreciated in conventional planning theory. This is so because planning is a basic human property with the implication that each agent operating in the city (person, family, company) is a planner on a certain level. In certain cases, because of the nonlinearities that typify the complexity of cities, the planned action of a single individual might influence the city more than that of the official planners and their plans. Urban dynamics can thus be seen as an on-going interaction between planners and their plans when none of them



Self-Organization and the City, Figure 6

a Burgess concentric zone model. (1) Central Business District, (2) Zone in transition, (3) Zone of working men's homes, (4) Residential zone, (5) commuters' zone. b Regional differentiation of Chicago as commonly presented in several geographical textbooks. Source: [95]

can fully determine the final form and structure of the city. They are all *participants* in a big city-planning game (see Part III in [95]).

The Inconsistency Between Self-Organization and Planning

All theories of cities were associated with city planning. The basic idea is that the city is an artifact and as such a product of humans' intentions and needs. Planning is needed in order to implement human needs and intentions in a rational way. The notion of complexity suggests that the city is a product of self-organization; if this is so, who needs city planning? The answer to this inconsistency has already been given above: the plans produced by city planners, like those produced by "ordinary" urban agents, are *participants* in a big city-planning game.

Complexity Theories of Cities – an Overview

The discussion in this section proceeds under the titles of eight "cities" that are related to general theories or specific methodologies. It starts with 'dissipative cities' to indicate that Prigogine's was the first complexity theory applied to the study of cities

Dissipative Cities

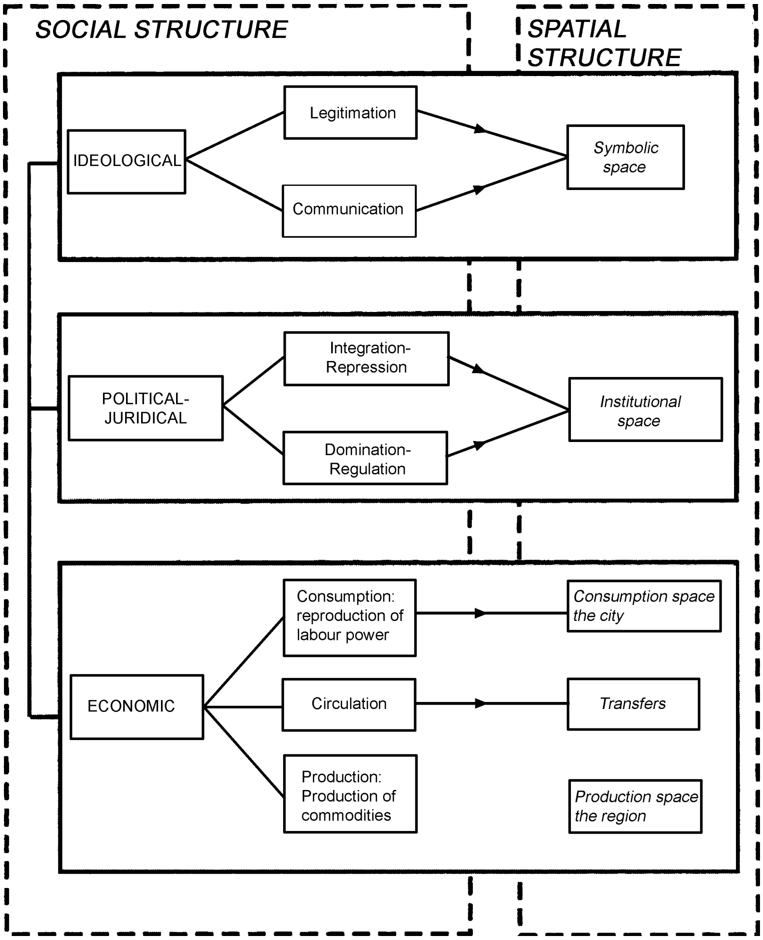
In their introduction to *Self-Organization in Nonequilibrium Systems* Nicolis and Prigogine [87] use the example of a city as a metaphor to convey to their fellow physicists what they mean by "self-organization."

> "An appropriate illustration would be a town that can only survive as long as it is a center for inflow of food, fuel … and sends out products and wastes."

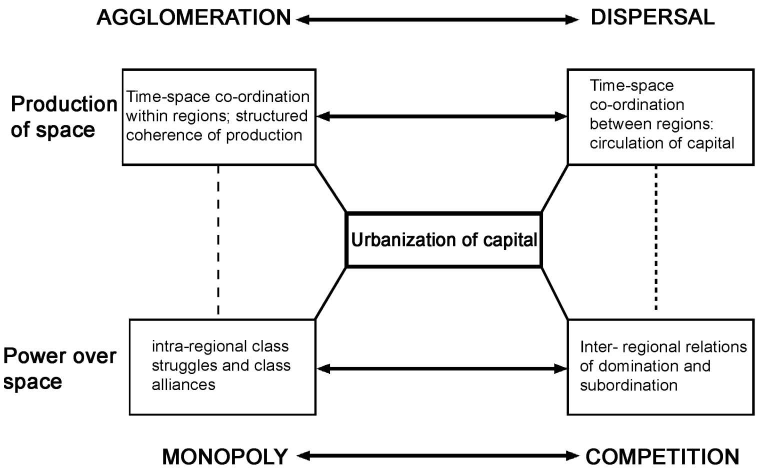
Peter Allen [5] – Prigogine's student – showed that towns and cities are not just metaphors, but genuine self-organizing systems. He did so by reformulating *central place theory* (above Sect. "Images of Cities") in terms of Prigogine's theory. (Note the resemblance between the hexagonal landscapes of *central place theory* and the hexagonal Bénard cells – one of the canonical experiments of the paradigm of self-organization).

Allen and co-workers' have developed a sequence of several models which elaborated their theoretical treatment of hierarchical landscapes of central places, first with respect to systems of cities in a given region and later at the intra-urban scale in connection with a single city. At a later stage they have also applied their models to real case studies of Brussels and the Belgian provinces [121], see also [109].

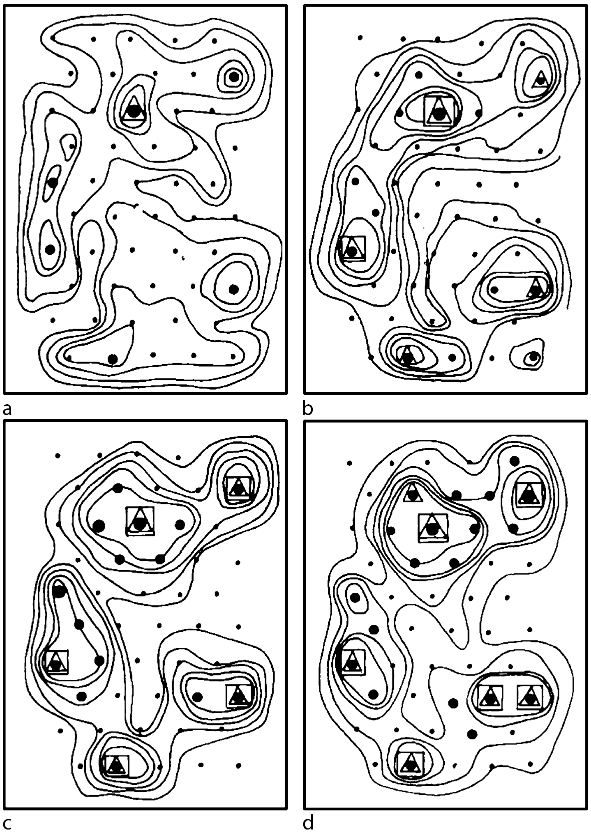
A typical model of Allan's starts with an infrastructure of localities in a region, each with its residents and jobs. The actors are individuals who migrate in order to get employment, and employers who offer or take away



Self-Organization and the City, Figure 7
The Marxist city as a spatial representation of social structure [33]. Source: [47]


AGGLOMERATION
DISPERSAL
Production of space
Time-space co-ordination within regions; structured coherence of production
Time-space co-ordination between regions: circulation of capital
Urbanization of capital
Power over space
intra-regional class struggles and class alliances
Inter- regional relations of domination and subordination
MONOPOLY
COMPETITION

Self-Organization and the City, Figure 8
The Marxist city as an outcome of basic tensions in the landscape of capitalism [63]. Source: [47]



Self-Organization and the City, Figure 9
Allen and Sanglier's simulated evolution of a dissipative system of cities. a at time (*t*) $t = 4$; b at $t = 12$; c at $t = 20$; d at $t = 34$

jobs depending on the market's situation. The migration/interaction between localities and the introduction and extraction of economic activities (i. e. employment opportunities), create for each locality a kind of local "carrying capacity" and for the system as a whole nonlinearities and feedback loops which link population growth and manufacturing activities. An example for a simulated scenario produced by the model is Fig. 9. It starts (Fig. 9a) with a hypothetical region characterized by a rectangular lattice of homogeneous localities. Then, the mere play of chance factors, such as the place and time where different enterprises and migrations start, produce symmetry breakings which entail an uneven distribution of population, employment and so on (Figs. 9b–d). The result is an evolutionary process by which new urban centers emerge, grow, and form the whole of the regional system of central places; as the system evolves, some old localities grow, others decline or even disappear, thus constructing the specific history of this region.

Allen and co-workers' approach exposes the similarity and difference between the "old" static approaches of Christaller and Lösch, and the new treatment by means of self-organization. In both, economic activities and interactions give rise to cities as central places. However, while in the old formulations the landscape reflects an equilibrium state which is the optimized sum of the properties of the various economic forces, the new landscape reflects a far-from-equilibrium situation in which the spatial hierarchical order among the central places is obtained, maintained and then transformed, by means of interplay between interaction, fluctuations and dissipation.

Synergetic Cities

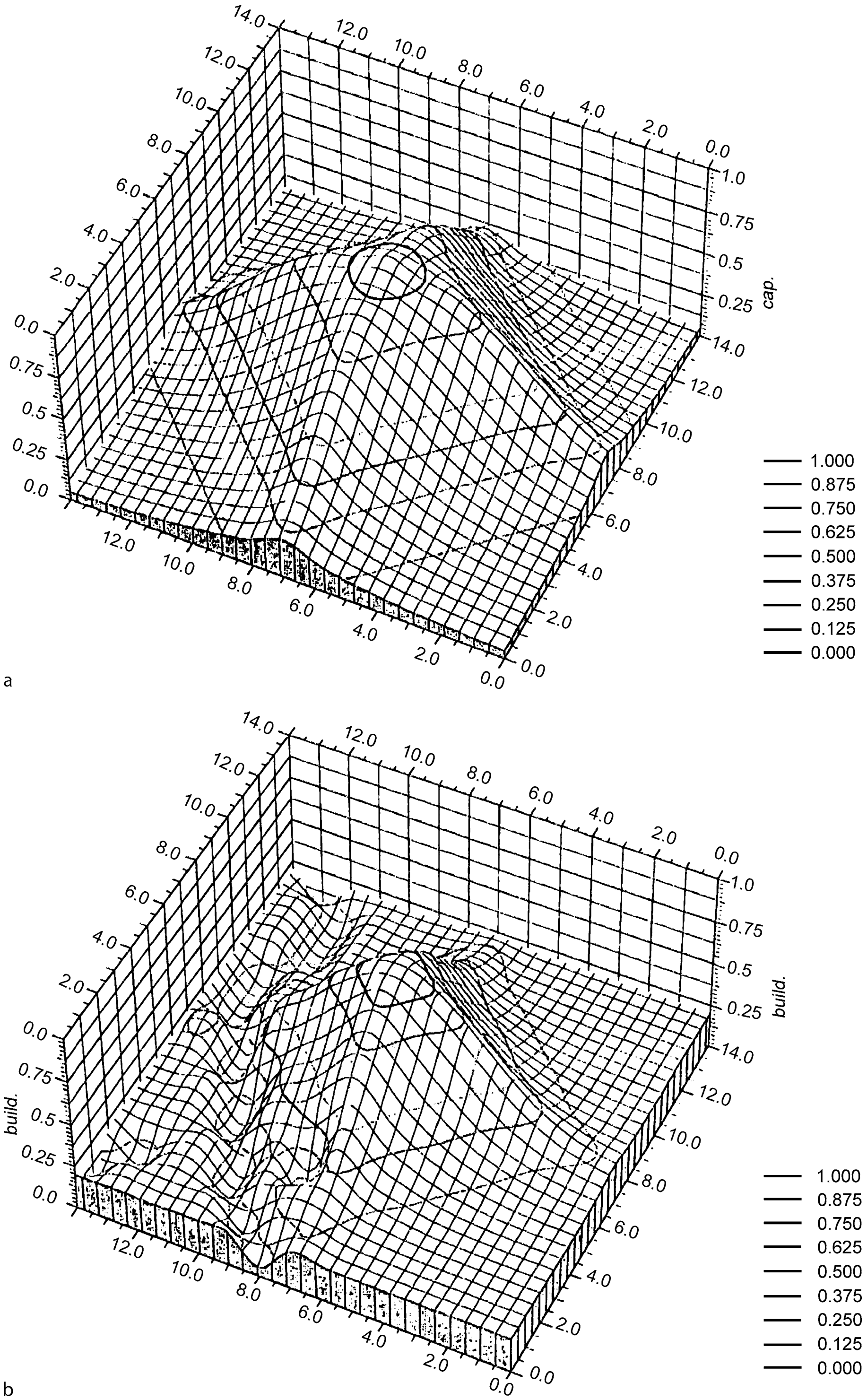
Two main approaches of synergetics have been applied to the study of cities. The first is the master-equation approach that is characteristic of Weidlich and co-workers studies in sociology, economics and urban dynamics [131,132,133,134]. For many years this was the main synergetic approach to cities, and most applications thus far have been within this conceptual frame [48,112,119]. The second, the pattern recognition approach, typifies the synergetics' treatment of pattern formation, cognition, pattern recognition and brain activities, as developed in the last three decades by Haken and co-workers [56]. Since the 1990s this approach has been applied to the study of cities as self-organizing systems [58,95].

Slow Cities and Fast Regions One way to look at Haken's synergetics and its slaving principle is in terms of interplay between slow and fast processes:

> If in a system of nonlinear equations of motion for many variables these variables can be separated into slow ones and fast ones, a few of the slow variables … are predestined to become "order parameters" dominating the dynamics of the whole system on the macro-scale [134].

This perspective stands at the basis of Weidlich's and co-workers studies on sociodynamics and cities [131,132, 133,134]. According to this perspective, fast and slow processes are easily identifiable in processes of settlement and urbanism. The fast ones typify the local microlevel of building sites, streets, subways, etc., whereas the slow processes typify the macrolevel of whole regions which are often described as systems of cities. The relations between the slow and the fast processes are described by the slaving principle: on the one hand, the regional system

> serves as the environment and the boundary condition under which each local urban microstructure evolves. On the other hand, the … regional macrostructure is … the global resultant of many local structures [134].



Self-Organization and the City, Figure 10

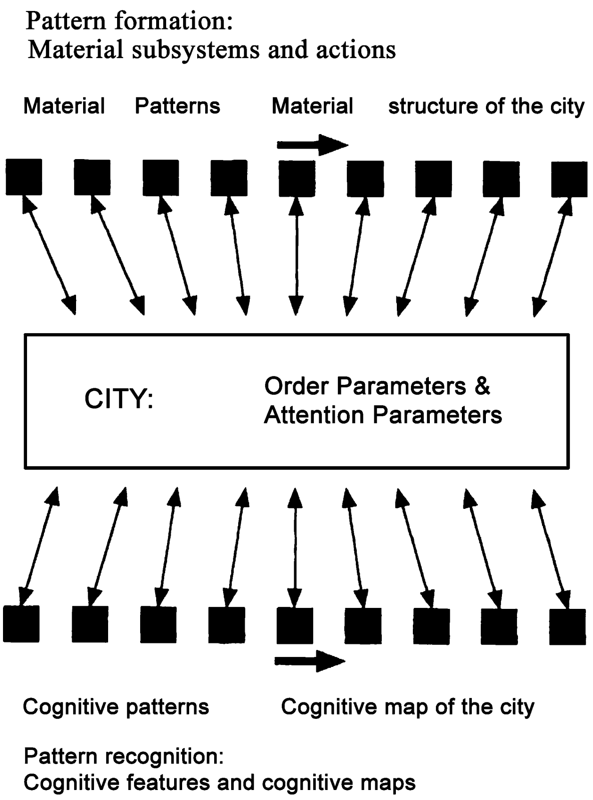
Building and development under population pressure [134]. ***Top*****: on a uniform urban plain.** ***Bottom:*** **on an urban plain with disturbances**

This circular causality between the local and the global, allows one to study global regional systems by assuming that local processes adapt to the slow regional ones, and to study local urban processes by treating the regional context as given, and of course to study the complex interplay between the local and the global. In all three cases Weidlich has recently prescribed a four stages approach: stage 1 concerns the configuration space of the variables; stage 2, measures the utility of each configuration; stage 3, defines transition rates between configurations which are in fact utility differences; stage 4 derives stochastic or quasi-deterministic evolution equations for the system under consideration. The central evolution equation is the master equation which defines the probability that the configuration under examination is realized at a certain time.

The above theoretical procedure has been used to study the role of population pressure in "fast and slow processes in the evolution of urban and regional settlement structures". Figure 10 brings some results from these studies, in which the city capacity for building and development is related to population pressure. Figure 10a shows the evolving city capacity when the urban plain is uniform, and Fig. 10b, when it is disturbed in one of its sites.

Pattern Formation and Pattern Recognition in the City

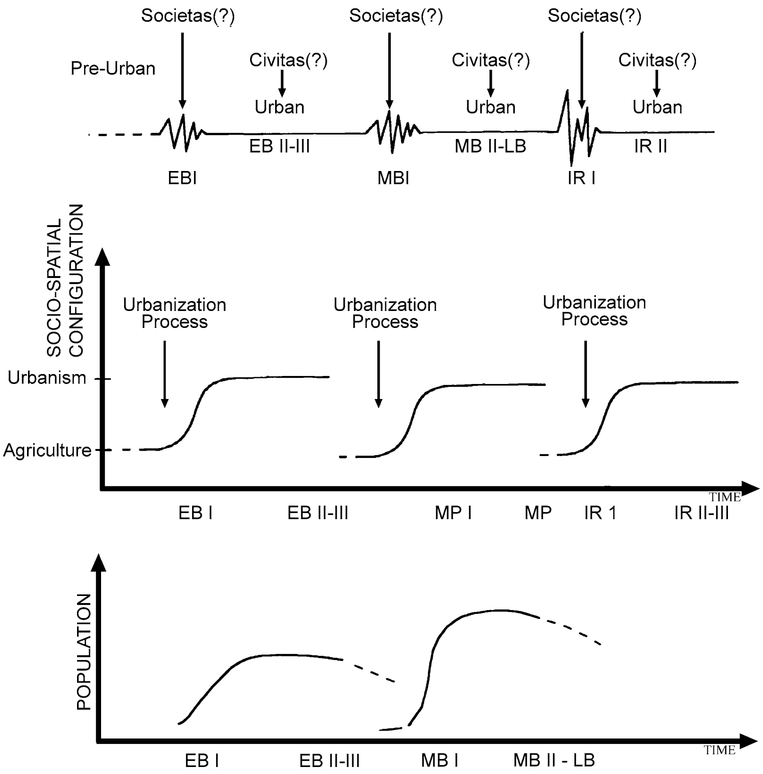
The paradigm of pattern recognition was derived by an analogy to the material process of pattern formation [54]. Haken and Portugali suggested that the synergetic pattern recognition paradigm is specifically attractive for the study of cities [58]. The latter can be perceived as self-organizing systems which are both physical and cognitive: individuals' cognitive maps determine their location and actions in the city, and thus the physical structure of the city and the latter simultaneously affects individuals' cognitive maps of the city. In their preliminary mathematical model Haken and Portugali construct the city as a hilly landscape which is evolving, changing and moving as a consequence of the movement and actions of individuals (firms etc.). The latter give rise to the order parameters which compete and enslave the individual parts of the system and thus determine the structure of the city. The new feature of this exposition is that the order parameters enslave and thus determine two patterns (Fig. 11): one is the material pattern of the city, and the other is the cognitive pattern of the city – its cognitive map(s).



Self-Organization and the City, Figure 11
The city as self-organizing systems which is at the same time both physical and cognitive. Its emerging order- and attention-parameters enslave the city's cognitive and material patterns. Source: [95]

Chaotic Cities

Self-organization is often regarded as a theory about *Order out of Chaos* [110] and yet, with a few exceptions [39,95] chaotic behavior is rarely studied in cities. Most complexity studies of cities perceive cities as ordered structures which the theory of complexity explains just how their order state was created. According to Batty (see p. 29 in [16]) this is "because the required growth rates [for chaotic behavior to appear in cities] are far too large". My view is that this is due, firstly, to the tendency of most students of complexity to focus on the short-term dynamics of Western cities from which perspective cities are indeed structurally stable. However, when the focus of interest turns to the long-term rural urban migration process in a country such as China, or to the archaeological record of the rise and fall of urban cultures [95,105], chaos suddenly appears. Looking at this *longue durée* [26] of cities, their evolution exhibits a very distinct and routinized path: a long period of "steady state", followed by a short period of strong fluctuations or chaos, from which the system re-emerges to a new level of steady state and structural stability, and so on (Fig. 12). As can be seen, the urban system moves from one structurally stable state to another, via bifurca-



Self-Organization and the City, Figure 12
The evolution of the settlement system in Palestine from the Early Bronze Age period to the Iron Age. Source: [100]. *Top*: A description in terms of chaos and order. *Middle*: A description of the process as a rhythm between agriculture and urbanism, interrupted by global collapses of the urban system. *Bottom*: Calculated population changes in the Early Bronze and Middle Bronze periods. Source: [93,95]

tions, when every evolutionary move is a transition from a microscopic chaotic state to an ordered, macro, steady state.

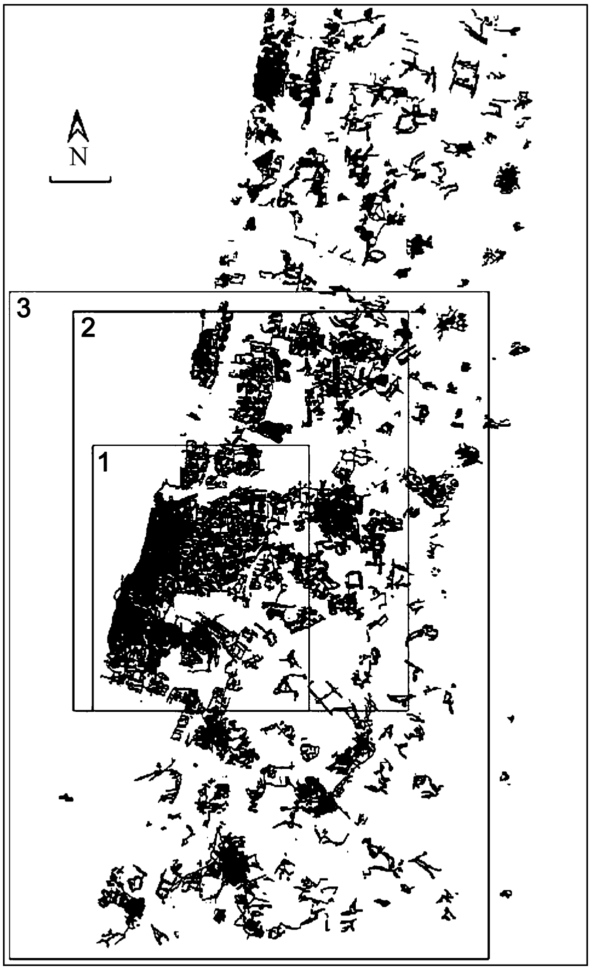
Secondly, this is due to the fact that phenomena of chaos and their role in cities during their short-term structurally stable periods have not as yet been fully studied. In a preliminary attempt to do so it has been found that often, when the city as a whole evolves stably, a few local unstable chaotic areas are found captive within the otherwise stable city. This phenomenon has been termed *the captivity principle* with the suggestion that it might play a supplementary role to Haken's slaving principle (see Chap. 5.8 in [95], [61]), namely, that these local islands of instability are needed in order to maintain the overall global stability of the city. Figure 15 below provides a hypothetical example simulated by means of cellular automata.

The play between chaos and order might show up also in the daily routines of cities. The movement of cars on the roads, or of pedestrians on pavements, are characterized by shifts between instable and stable motions and might thus be candidates for this kind of interpretation.

Fractal Cities

Mandelbrot's theory of fractals is based on the notions of *self similarity* and the *fractal dimension*, and, on the idea that a rather simple iterative process might produce highly complex geometrical shapes. Using these principles, several scholars have demonstrated, first, that the complex geometries of urban form, growth and evolution, on intra-urban and inter-urban regional scales, can be generated by means of a simple iterative process with a few and simple rules. Second, that many urban structures are self-similar and have fractal structure. The most comprehensive study in this domain is Batty and Longley's [19] *Fractal Cities*, to which one can add studies on urban structure, on the fractal structure of transportation networks [23], on the question "when and where is a city fractal?" [24] and more (For updated survey of studies see [16]). Figure 13 illustrates the evolving fractal structure of the Tel-Aviv metropolitan area from 1935 onwards.

Another important insight implied by fractal cities studies is that a city, or a system of cities, in a steady state

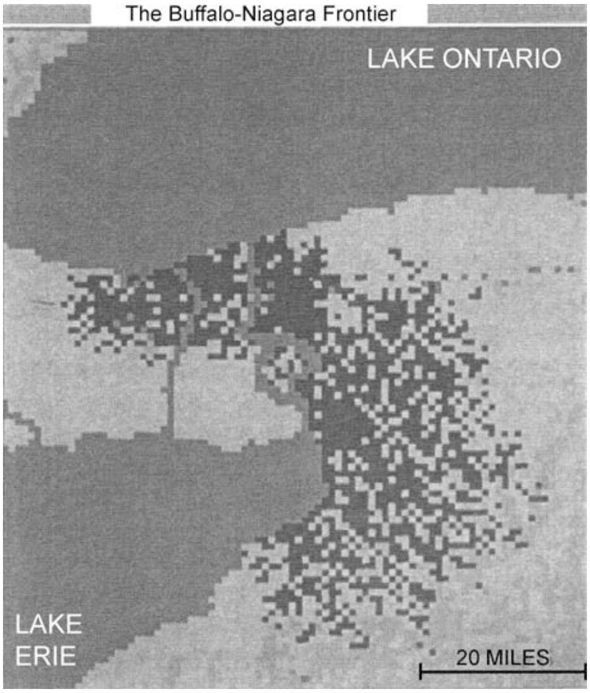


Self-Organization and the City, Figure 13
The evolving fractal structure of the Tel-Aviv metropolitan area from 1935 onwards: The central parts 1 and 2 were fractal during the entire period, while their fractal dimension increased with time. The entire metropolitan area became fractal only after 1985. In 1991 the fractal dimension of the Tel Aviv metropolitan area was found to be 1.667 with error of 0.037. Source: [23]

does not mean equilibrium and stability, as is the case of Christaller's and Lösch's central place theories, for example, but rather a rich and complex evolution and change according to a given ordering principle.

Cellular Automata Cities

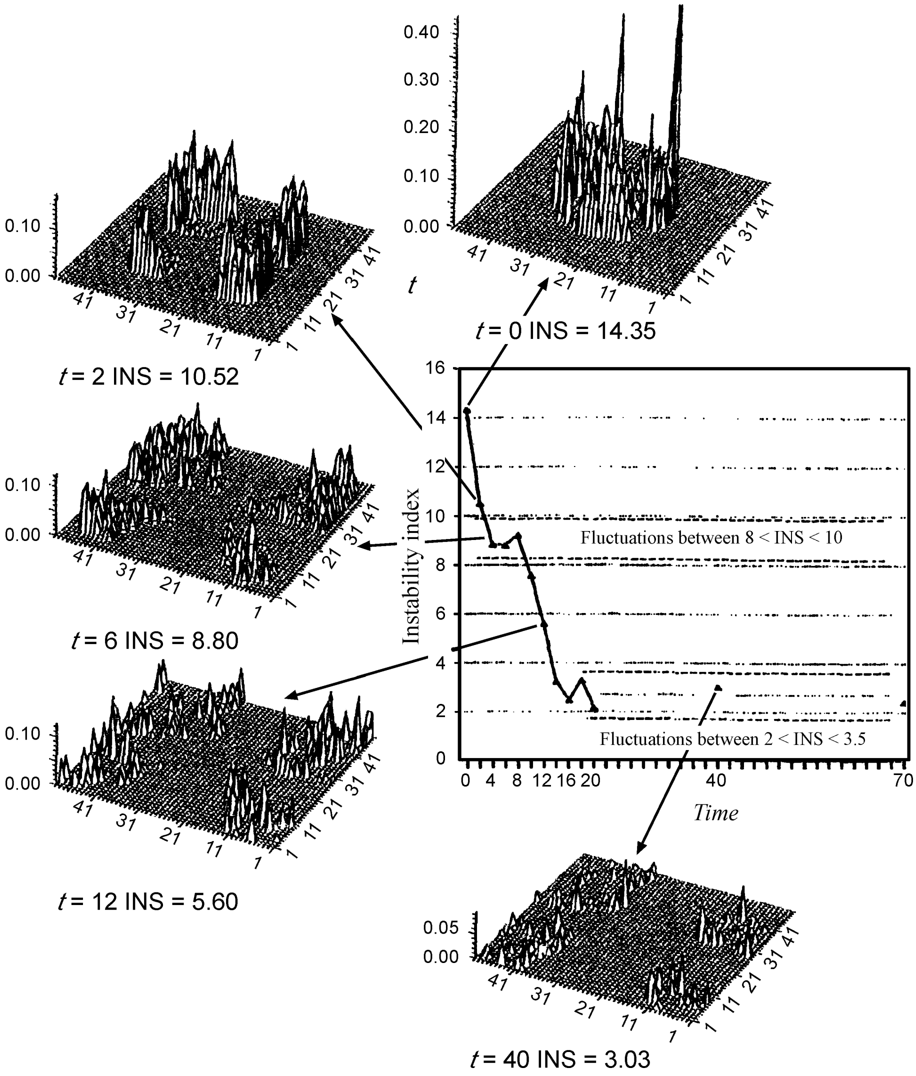
The attraction of CA (cellular automata) models to the study of cities is almost self-evident. Real cities are built of discrete spatial units such as houses, lots, city-blocks and the like. CA models are also built of discrete spatial units – the cells. In real cities the properties of local spatial units (e. g. land value) are determined, to a large extent, in relation to their immediate neighbors; so are the properties of the cells in CA models. These resemblances, together with the mathematical simplicity of CA models, make them natural tools to simulate urban processes. In the last few years CA urban simulation models are among the most popular approach to simulate the dynamic of cities [16,22,95].



Self-Organization and the City, Figure 14
Cellular automata simulation of the Buffalo–Niagra frontier. Source: [17]

One can divide the various models of CA cities into implicit and explicit self-organized CA cities. The first group refers to studies the aim of which is to use the simulation capabilities of the CA city in order to best-fit a certain simulated pattern to an existing one [16]. Figure 14 is an example. The general motivation here is to explain an existing or historical pattern, or alternatively to predict a future one for the purpose of planning. The fact that the model has properties of self-organization, just adds more realism and sophistication to the simulation.

The second group concerns explicit self-organized CA cities. Here the central motivation is to use the model as means to investigate the self-organization properties inherent in cities and urbanism. For example, how micro decisions and behavior of individuals and firms, taken at the local scale, are related to the global behavior and structure of the city. Such models are essentially heuristic and they regard the simulated CA city as essentially a learning device (Fig. 15).



Self-Organization and the City, Figure 15

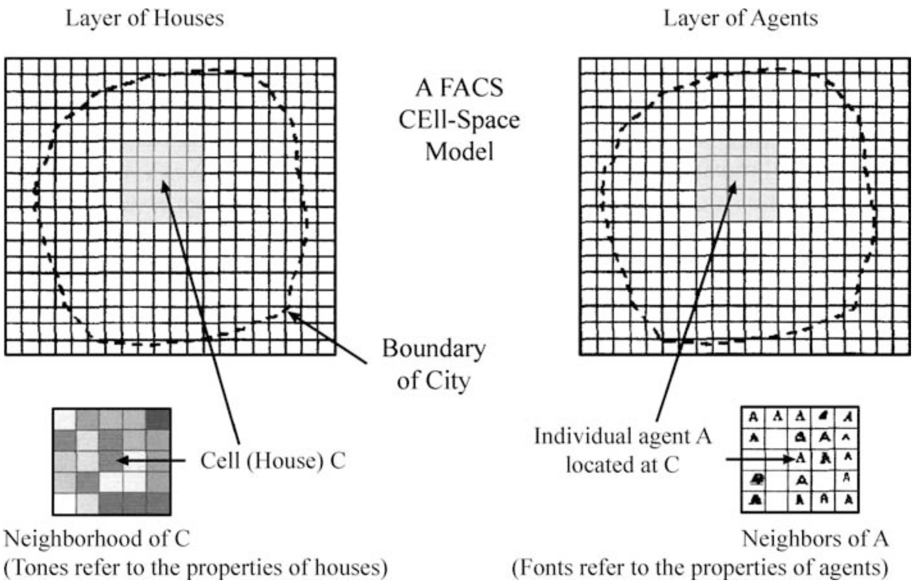
Time evolution of consecutive stages of SIS (stability-instability surface) in the development of a city with 33% neutral Greens when the rest of the Greens and all the Blues are segregatives. Source: [95]

Because of their iterative structure, CA models can be used as convenient tools to generate fractal structures [16,17] and the insight they add to our understanding of cities is similar: an iterative process guided by a few simple rules, can generate complex structures such as cities [16].

AB and FACS Cities

CA is an efficient tool to model the relations between infrastructure objects of the city. Unlike infrastructure object, urban agents have aims and plans, can learn and move in the city, and see and know beyond their nearest neighbors. *Agent base* (AB) models that are built to imitate such cognitive entities were applied to the city dynamic too. An important source of inspiration here was Schelling's model that demonstrated how local and simple behavior of urban agents can give rise to complex residential segregation in cities – even when their tendency for segregation is minimal [122]. Subsequent agent base studies have supported Schelling's finding [16] and added that a small minority of agents with a tendency for segregation is sufficient to turn the whole city into a segregative structure [95].

Free agents on a cellular space (FACS) models combine CA and AB models [95]. A typical such model is built as a superposition of a CA layer simulating the relationship



Self-Organization and the City, Figure 16
A typical FACS model is constructed of two-layers: a population layer of human agents describing the migratory and interaction activities of individuals (*right*), superimposed on a CA infrastructure describing the urban landscape (*left*). Source: [95]

between the city's infrastructure objects (buildings, roads …) and AB layer that simulates the activities of the urban agents (Fig. 16). At each model iteration new agent(s) come to the city with a certain intention in mind – say to find a house to live in. The agent then examines the available empty cells/buildings, ranks them according to its set of preferences and picks the best one. Once the agent located itself in a certain cell, the CA dynamics starts: The properties of each cell are determined by reference to the properties of its neighbors; and if the cell is occupied by a certain agent, by some mix between the properties of the agent and its neighbors. Figure 17 presents typical results.

Sandpile Cities

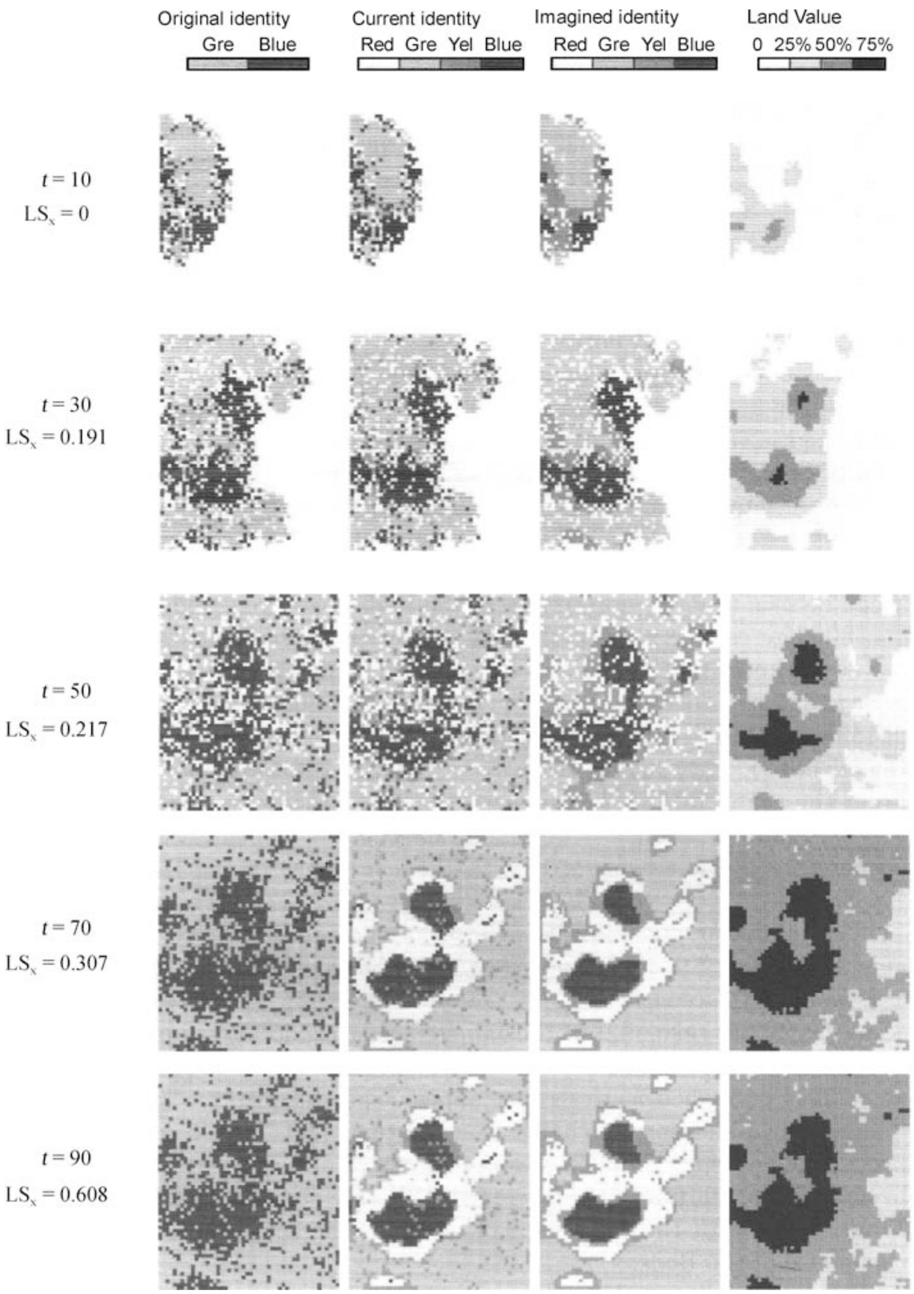
The sandpile, the canonical example of *self-organized criticality* [11,12,16,20], has two incongruous features: the system is unstable in many of its local locations; nevertheless its global state is absolutely robust: The local configurations of the sand change all the time because of the avalanches, while the statistical properties, such as the size distribution of the avalanches, remain essentially the same. Similarly to the sandpiles, cities appear volatile and fast moving at their local scales while at their global scale they appear absolutely robust [16,20]. For example, the size-distribution of many cities and systems of cities remains essentially the same under circumstances such as ongoing population growth (Above Sect. "Image of Cities").

Compared to the "grand" synergetic and dissipative cities, the sandpile city is a kind of a zooming-in to the internal dynamics of self-organized cities in their steady state periods – when they are controlled by what in synergetics is called order parameters. Sandpile cities show how complex and rich is the internal dynamics of a city in steady state (Fig. 18).

Small World Cities

The notion *network* is implicit in all theories of complexity. Recently, Watts and Strogatz [130] showed that complex networks have 'small world' characteristics [84] and Barabasi and Albert [14] demonstrated that complex networks are scale free thus following the *power law* that according to Barabasi [13] is a mark of self organization.

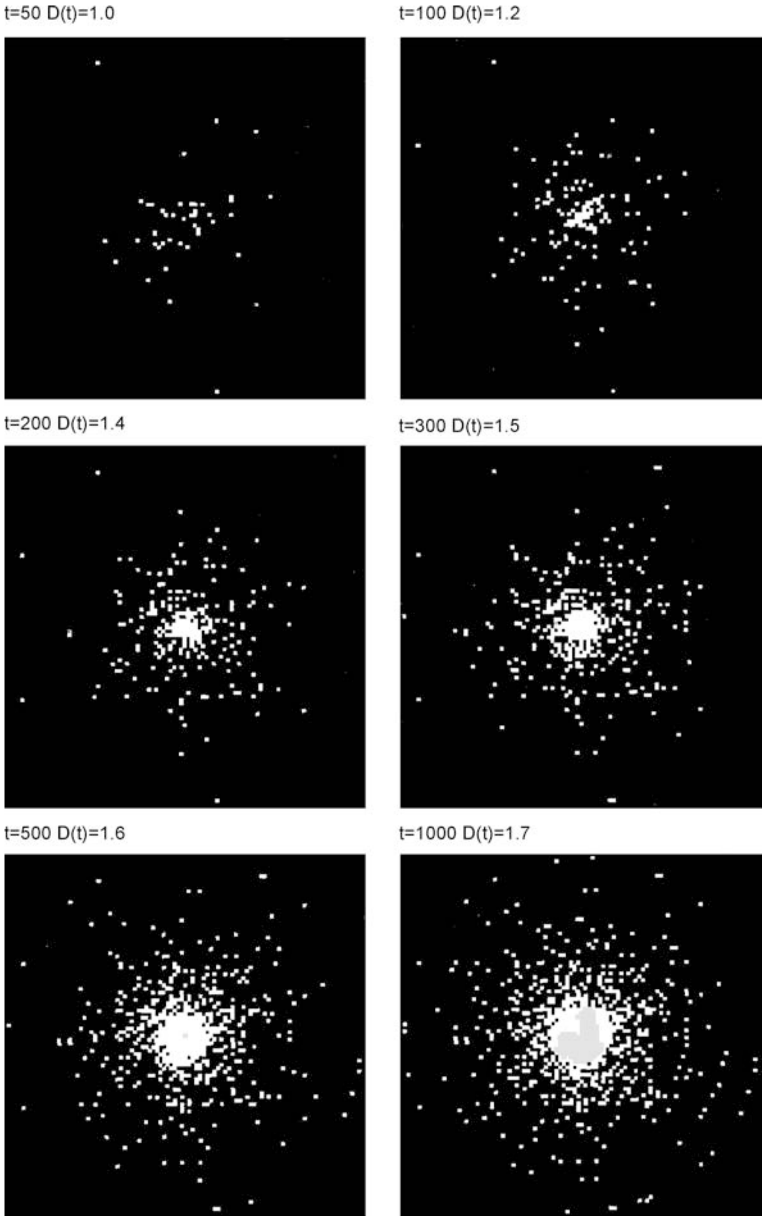
The link to cities as complex systems was just natural: The view of systems of cities as networks characterized by the power law was indicated above. Single cities too were described as networks. Thus Alexander's classic "a city is not a tree" [1] demonstrated that cities are typified not by a simple *tree network*, but by a complex *semi-lattice network* (Fig. 19). Alexander's view was recently reformulated in terms of the new science of networks [117,118]. Another example is Hillier's *space syntax* that analyzes the morphology of urban spaces in terms of networks [65,66]. Space syntax exposes the way society



Self-Organization and the City, Figure 17
Several snapshots from an evolving FACS city. Source: [95]

determines the urban morphology and the way the letter feeds back and re-shapes society. The link between space syntax and network analysis has already produced several useful results [38,41,67,89,114].

In the domain of transportation one can mention studies that characterize roads' traffic dynamics in terms scale-free networks (Fig. 20) [68,72], the same was found for the transit system in Beijing [137], pedestrian movement [73] and for the canal networks of Venice [25]. Andersson et al. [9] showed that the market dynamics generates land values that can be represented as a growing scale-free network. Finally, Batty [16] has suggested viewing cities and their dynamics from the combined perspectives of networks, fractals, self-organized criticality and AB.



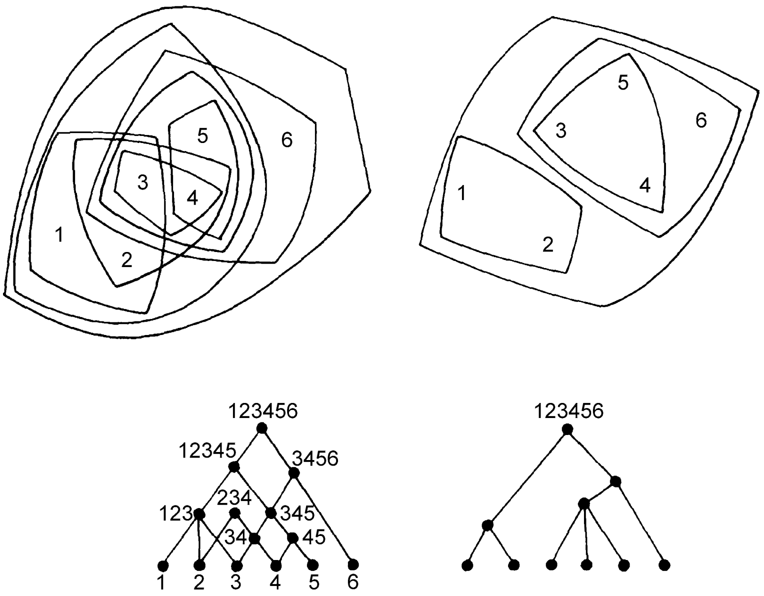
Self-Organization and the City, Figure 18
Self-organized criticality: "Simulation of a hypothetical urban growth pattern in its critical level". Source: [16]

Self-Organization and the City

Self-Organization and the City is an ongoing project that explores the city as a complex system from two interrelated perspectives: Haken's (1983) synergetic theory of complex systems, in particular from the perspective of the pattern recognition paradigm [50,51,55] and IRN – inter-representation nets [94]. The link between the two is termed SIRN – synergetic inter-representation nets.

SIRN – Synergetic Inter-representation Networks

IRN commences with a distinction between cognitively *simple tasks* that can be performed by working memory (e.g. $2 \times 3 = 6$) and *complicated tasks* (e.g. $257 \times 389 = 99\,973$) that are the result of the "magic number seven" that constraints our ability to process information in working memory [85]. One way to overcome this limitation is by means of IRN: We first externalize the task



Self-Organization and the City, Figure 19
A tree network (*right*) versus a semi-latice network (*left*). Source: [1]

(write it down on a paper); then we solve part of it internally ($9 \times 7 = 63$); externalize it again and so on in a sequence until the task is completed.

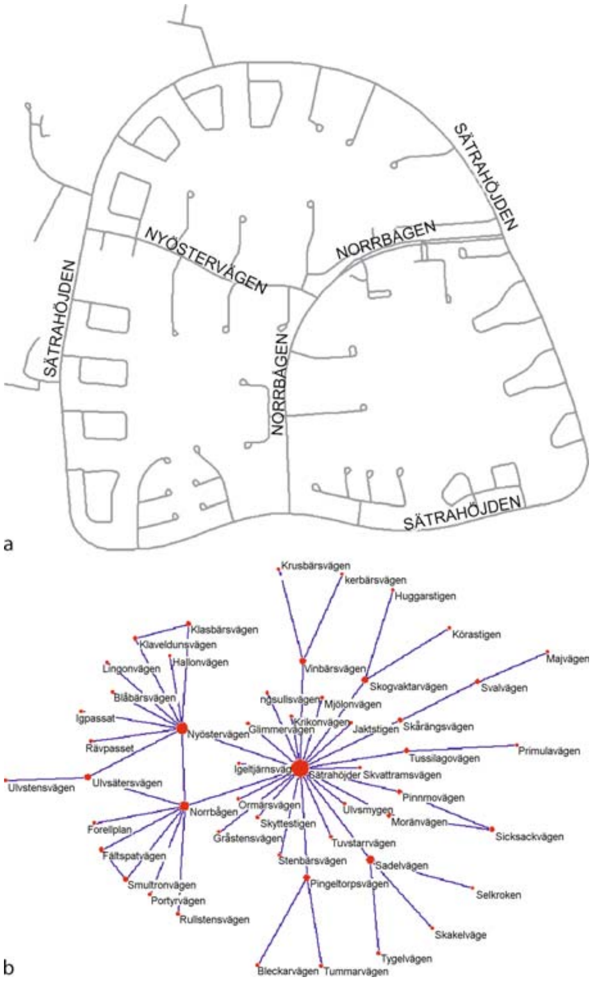
Complex tasks refer to creative cognitive tasks, when a person writes, paints, designs etc. Such a task often starts with a vague idea in mind that the person then externalizes by writing it down or painting … Here too the process proceeds by interplay between internal and external representations, but with one important addition – it involves emerging properties. It is here were synergetics gets in and the process becomes SIRN. More specifically, the process might start with a preliminary internal idea (or external cue that entails internal idea), that the person then externalizes and so on. After a few internal-external iterations an order parameter (in the sense of synergetics) emerges and enslaves subsequent iterations.

The development of the notion of SIRN was inspired by Bartlett's serial reproduction scenarios in his study *Remembering* [15]. A typical such scenario starts when a test person is shown a text or a figure and is asked to reproduce it out of memory (Fig. 21). The result is offered to a second person that is asked to do the same and so on. As shown by Bartlett, at the beginning the reproductions change from person to person, however, at certain stage they stabilize and become a *scheme*. Stadler and co-workers [126] demonstrated that the scenarios proceed as synergetic self-organized process. The focus of interest in the above studies was on the way schemata are created. Haken and Portugali have used the Bartlett scenarios as illustration of the play between internal and external representations [59,94].

It is important to emphasis, first, that external representations are media that enable communication between persons and the emergence of collective SIRNs – e. g. a brain storm. Second, that internal and external representations are *generative* – once produced, they generate new ideas and properties not seen before in previous representations.

The Basic SIRN Model Haken and Portugali [59] have cast the SIRN process into the formalism of synergetics. They started with Haken's [54] 'synergetic computer' (Fig. 22, *Top*), composed as it is of an input layer with model neurons representing the initially given input activity; a middle layer representing the order parameters, and an output layer with neurons representing the final activity of each neuron. The first step is to look at this network from the side, as indicated by the arrow. The result is shown in Fig. 22, *bottom, left*. Adding to the latter external inputs and outputs, we arrive at our basic SIRN model (Fig. 22, *bottom, right*) that has two kinds of inputs, internal and external and two kinds of outputs, again internal and external. The middle node symbolizes the order parameters that emerge out of the interaction between internal and external representations.

The basic SIRN model can be seen as symbolizing a self-organizing active agent that is subject to two flows of information: internal and external (Fig. 23). The first



Self-Organization and the City, Figure 20

A small street network (a) and its connectivity graph (b). Every node in b is labeled by the corresponding street name, and the size of nodes shows the degree of connectivity of individual streets. Source: [73]

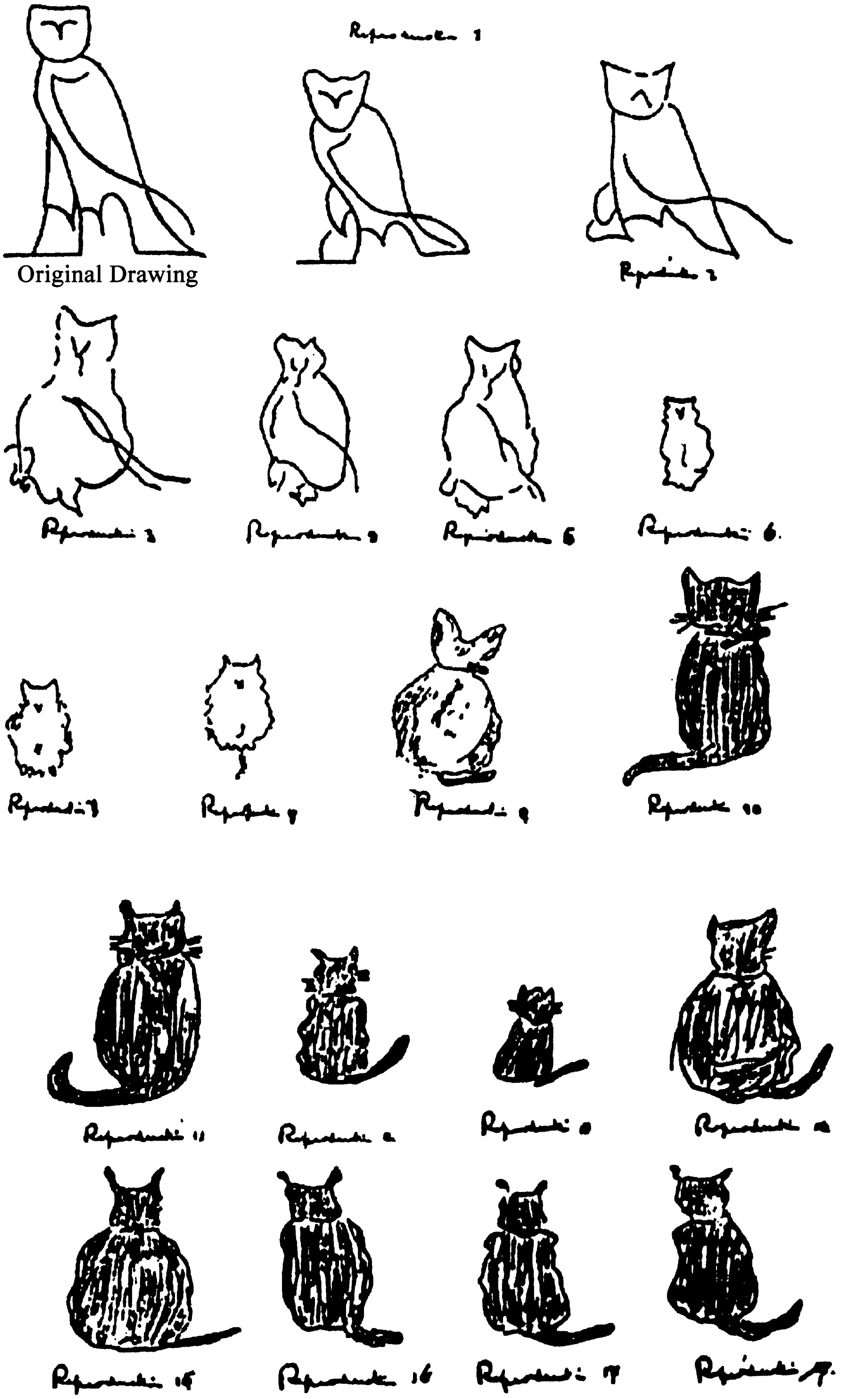
is coming from the mind/brain, in the form of ideas, fantasies, dreams, thoughts, and the like, while the second from the 'world' – via the senses, the agent's body and/or artifacts. The interaction between these two flows gives rise to an order parameter that governs the agent's action and behavior, as well as the feedback information flow to the agent's mind. 'Action or behavior' may refer to a single individual executing exploratory behavior, reproducing texts or drawing, as well as to several individuals collectively reproducing a large-scale artifact such as a city. In an analogous fashion, the 'feedback information flow' refers to the formation of internal representations, such as images or learned patterns. The order parameters are determined by a competition in line with the synergetics' pattern recognition paradigm noted above. Note that all the above steps (and below), can and have been, performed by a computer so that the approach is entirely operational.

In order to apply the basic SIRN model to specific case studies, Haken and Portugali [59] reformulated it in terms of three prototype sub-models: the *intrapersonal*, the *interpersonal collective*, and the *interpersonal with a common reservoir* sub-models (Fig. 24). The first refers to a solitary agent, the second to a sequential dynamics of several agents whereas the third to a simultaneous interaction. The third sub-model is, in fact, a theory of urban dynamic. The intrapersonal is typical to the way of an artist, for instance, develops her/his work (Fig. 25), whereas the interpersonal to the Bartlett scenario that provided a source of inspiration to IRN (above, Fig. 21).

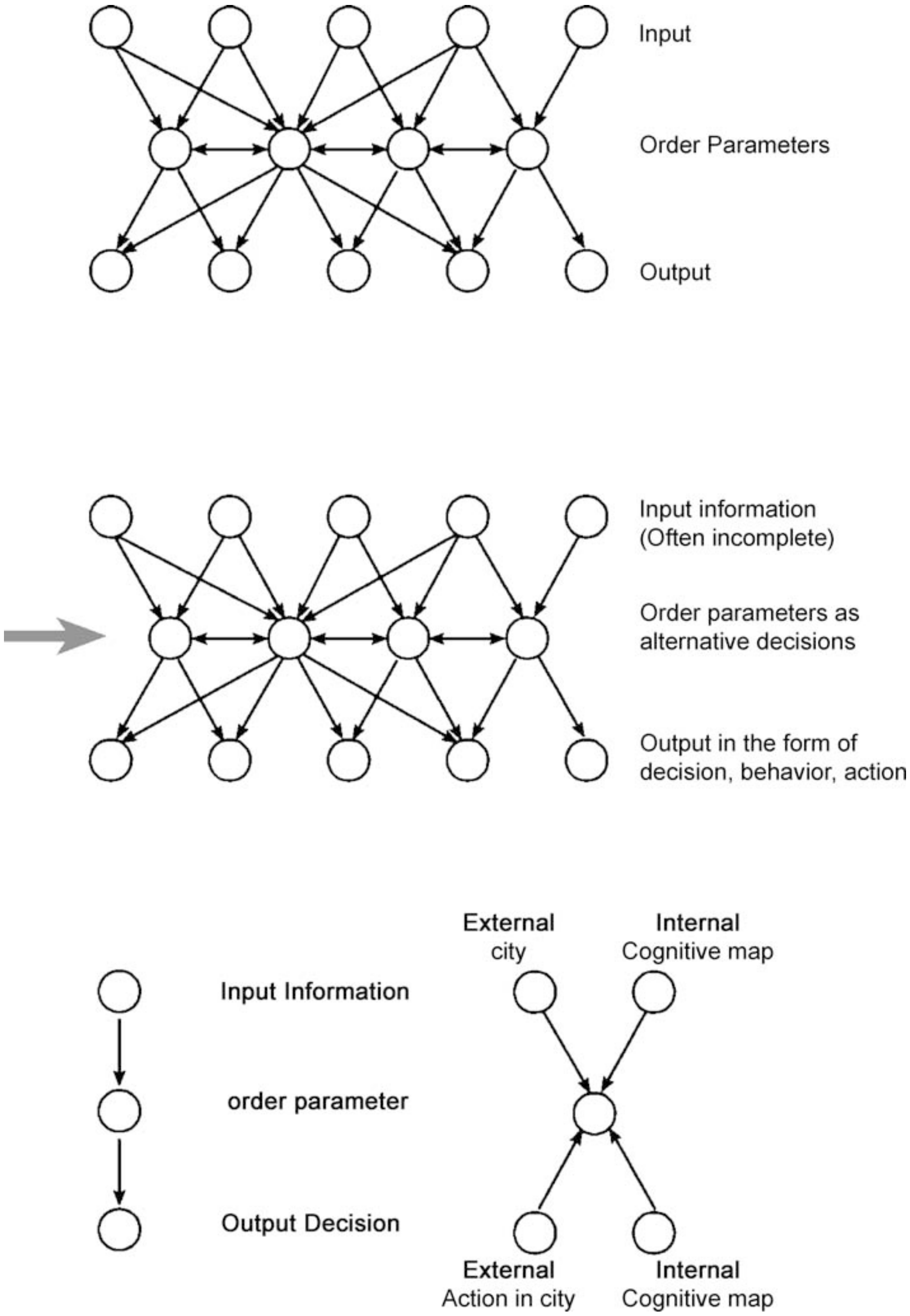
In the first two sub-models the process depends fully on the biological memories of individuals. In the third sub-model the process depends partly on biological memories, as before, but partly also on externalized non-biological memory that we term a *common reservoir*. This common reservoir of external, artificial and non-biological memory, might take the form of texts, Internet, buildings or whole cities.

Figures 24 (*bottom*) and 26 illustrate graphically this public-collective SIRN sub-model. Each individual agent is subject to internal input constructed by the mind/brain, and external input which is the legible information coming from the common reservoir, that is, the city. The interaction between these two forms of input gives rise to a competition between alternative decision rules that ends up when one or a few decision rules "wins". The winning rule(s) is/are the order parameter(s) that enslave(s) the system. The emerging order parameter governs an external output, which in the case of a city is the agent's behavior and action in the city, and an internal output, which is an information feedback loop back to the agent's mind/brain.

Both the previous sub-model and the present one involve a two-scale self-organization process: an individual-local scale referring to each individual agent as a self-organizing system, and a collective-global scale, referring to the whole city as a self-organizing system. The individual agents by their action and behavior determine the city, which by means of its emerging order parameter(s) enslaves the minds of the individual agents. In the language of synergetics this process is termed *circular causality*. In terms of social theory it is close to notions of socio-spatial *reproduction* and *structuration*. Recent applications show that the common reservoir might be a non-biological externalized memory such as a city[95,96,97,98], a planning



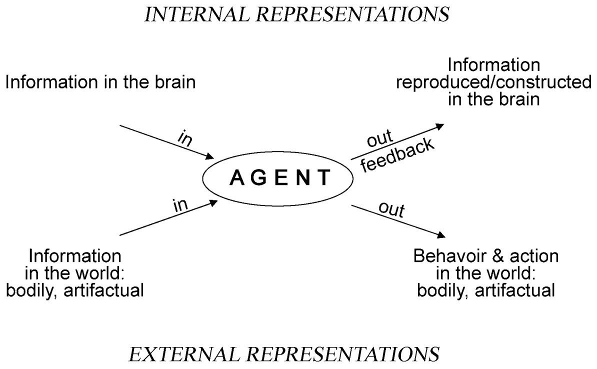
Self-Organization and the City, Figure 21
A Bartlett's scenario of serial reproduction: an Egyptian 'Mulak' (owl) transformed into a cat (see pp. 180–181 of [15])



Self-Organization and the City, Figure 22
The derivation of the SIRN model. See text

textual report or an urban planning policy emerging out of a discourse among the members of a planning team [105]. Note that as in the previous model, here too, due to circular causality, as the process evolves the subjective cognitive maps of the individual agents are becoming more similar to each other and an inter-subjective, collective cognitive map emerges. Both private–subjective cognitive maps and public-collective ones are thus *constructions*.

The City Game A simple and effective way to illustrate the SIRN view on the dynamics of cities is by means of a set of experiments termed *city games* [94]. A city game


INTERNAL REPRESENTATIONS
Information in the brain
Information reproduced/constructed in the brain
in
out
feedback
AGENT
in
out
Information in the world: bodily, artifactual
Behavoir & action in the world: bodily, artifactual
EXTERNAL REPRESENTATIONS

Self-Organization and the City, Figure 23
The SIRN model symbolizes a self-organizing agent that is subject to two forms of information: internal and external, and is actively constructing two forms of information, again internal and external. Graphically, Fig. 23 corresponds to Fig. 22 (*bottom, right*) turned 180° on its NW-SE axis

can be described as a group dynamics that involves some 40 to 70 participants. Their aim is to build a city on a floor, representing the site for a new city. Each player is given a 1:100 mock-up of a building and in his/her turn is asked to locate it in the virtual city on the floor, in what s/he considers as the best location for that building. In a typical game (Fig. 27a and Fig. 27b), the players observe the city as it develops, and in the process also learn the spontaneously emerging order on the ground. It is typical in such games that, after a few initial iterations, an observable urban order emerges. The participants internalize this emerging order and tend to locate their buildings in line with it. As can be seen, the main features of such a game are the main ingredients of SIRN, namely, a sequential interplay between internal and external representations, the emergence of a collective complex city as an artifact and a typical synergetic process of self-organization. Needless to say that the city game is not a 1:1 description of reality, but an illustration of the dynamics of cities as *dual* self-organizing systems.

Cognition and the City

SIRN is at once a theory of cognition, cognitive mapping and urban dynamics. This emphasis on cognition is a direct consequence of complexity theory; a major achievement of complexity theory was to show how local behavior and interaction between urban agents give rise to the global structure of the city. The agent is thus the main and most important actor. Given this, one would assume that practitioners of complexity theory and urban simulation models will have an elaborated theory of agents' perception, behavior, decision making and action; especially so, in light of the fact that a whole body of research on agents' behavior was readily available. I'm referring to studies on spatial cognition, spatial behavior and cognitive mapping that were developed on the interface between cognitive science and urban studies [45,77,92,98,99,103]. And yet, with few exceptions such as SIRN, this body of theoretical and empirical studies is largely overlooked by students of complexity theory of cities. Researchers in this field tend to follow economists by assuming that individuals behave in space as simple "economic persons". The result is that the rather *simple* behavior of agents in the models contradicts the *complex* behavior revealed by studies on cognitive mapping and spatial behavior.

In *The Sciences of the Artificial* Simon [124] suggested that the observed complex behavior of human agents, guided as it is by aims, plans, intentions, needs, policies and so on, misleads us as it is only an external appearance of innately simple behaving entities: Similarly to simple animals, we humans as

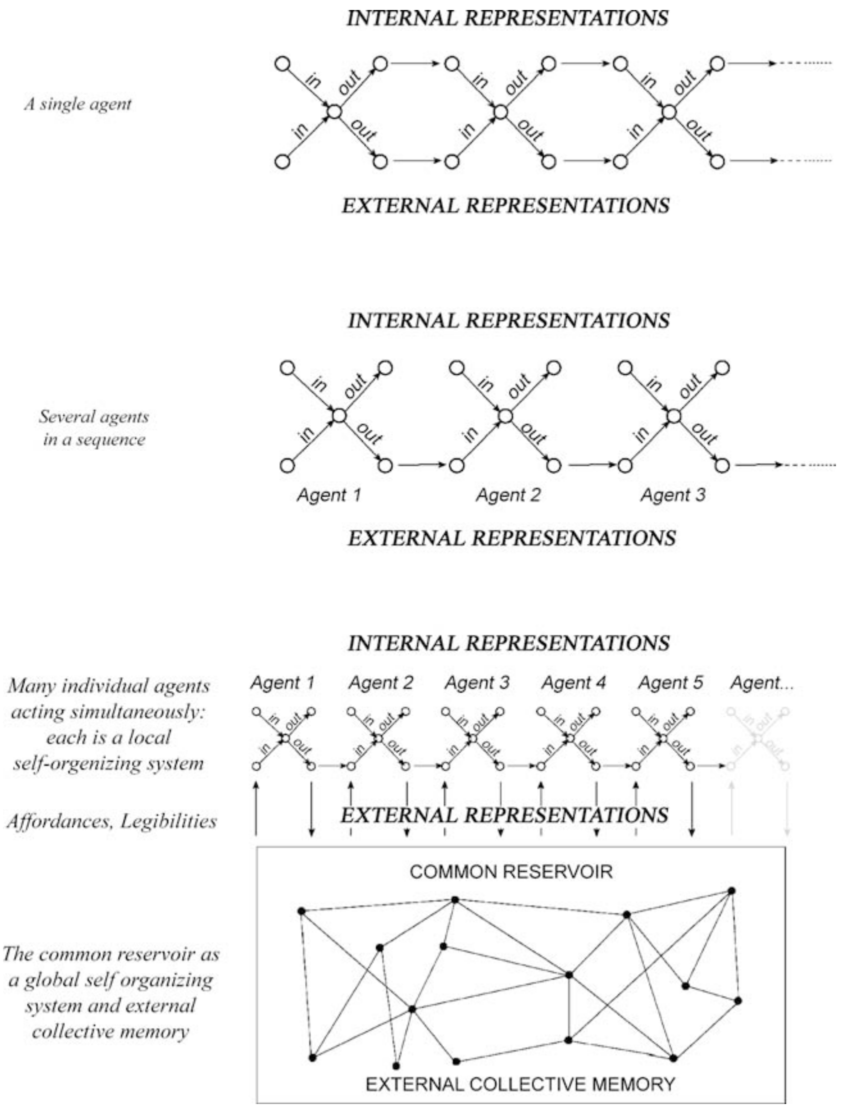
> "behaving systems, are quite simple. The apparent complexity of our behavior over time is largely a reflection of the complexity of the environment in which we find ourselves" (ibid. 53).

Most AB/CA urban simulation models are built in line with Simon's logic. They typically start with local interactions between agents having a few simple aim(s) "in mind". This interaction gives rise to an urban system, which from iteration to iteration becomes increasingly complex. Complexity is thus understood as a property of the whole global system, but not of its individual parts.

The efficiency of the *simple cause→complex effect* model is apparent. But there is a catch here: Several empirical studies, of animals' and humans' exploratory behavior, for example, falsify Simon's view [96,97,98]. Furthermore, the property of the city as a *dual* self-organizing system implies that the initial conditions of such complex systems are relatively large numbers of interacting parts, each of which is itself a complex system exhibiting complex behavior. Can there then be a science of cities that is not based on Simon's model? The answer is yes! To see how we shall look at the relations between self organization and information.

Information Compression, Inflation, Adaptation

Complexity is a property of systems that exchange matter and *information* with their environment and that their huge number of parts forms networks characterized by complex feedback and feedforward loops that allow intensive flow of *information* inside the system.

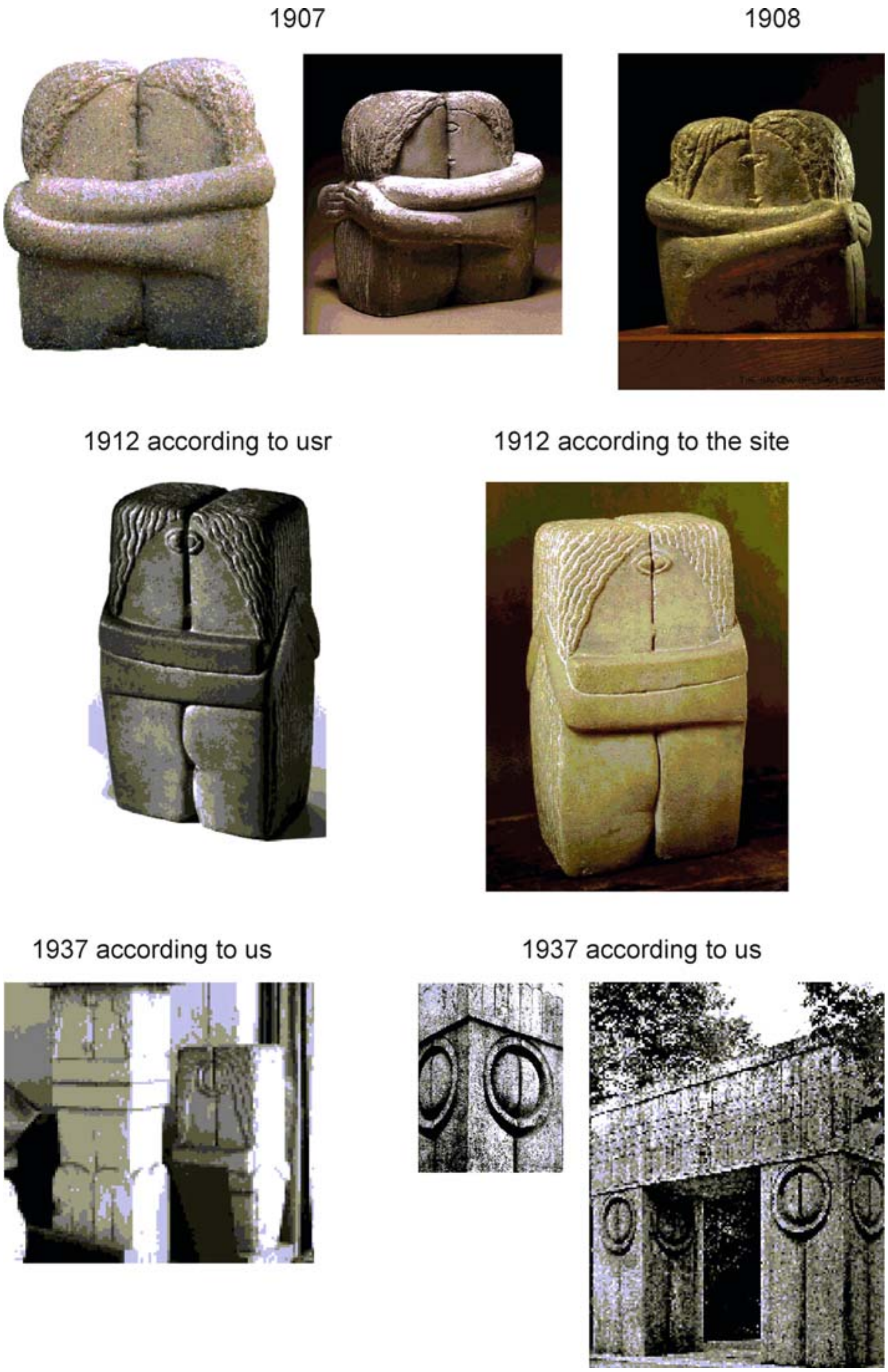


Self-Organization and the City, Figure 24

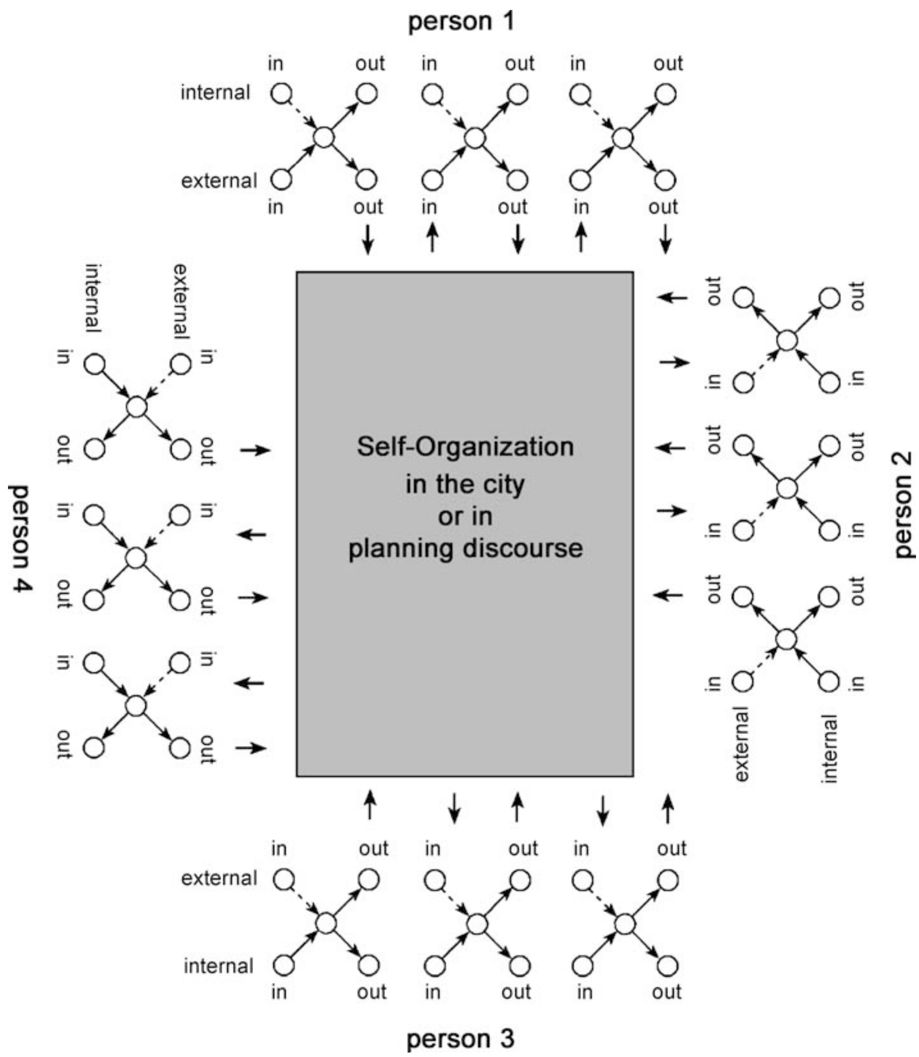
***Top:* The intrapersonal SIRN submodel of a single person. *Middle:* The interpersonal submodel: serial reproduction of several persons. *Bottom:* The interpersonal with a common reservoir submodel. Note that in the intrapersonal submodel information is transmitted via external and internal outputs, in the interpersonal via external output only (action and behavior), while in the third submodel information and interaction between the agents are mediated by the common reservoir (e. g. a text, a city, Internet, etc.). Source: [96]**

The notion *Information* is associated with Shannon's theory of information [123] that has played a seminal role in the development of system thinking. In Shannon's theory the notion of information is a pure quantity (usually measured by *bits*) devoid of any meaning. Such a concept of information makes sense only in closed systems where the number of possible states the system can take is finite and a-priori known; hence the link between information and the notion of Entropy, which is a property of closed systems. For example, the information conveyed by throwing a die is 2.5 bits, that is, the logarithm to the base of 2 of the six possible states the system can take. But the complex systems we are dealing with are by definition *open*. So what is the meaning of information in complex systems?

In *Information and Self-Organization* Haken [52] suggested that complex systems 'self-organize', that is, 'interpret', the information that comes from the environment. In other words, the meaning assigned to the message depends on the receiver (the receiving system) and not just on the message itself as in Shannon's theory. Haken (see



Self-Organization and the City, Figure 25
'The Kiss' by Brancusi: from a figurative kiss in 1907, to the geometrical 'Gate of the Kiss' in 1937: An intrapersonal SIRN process in sculpturing

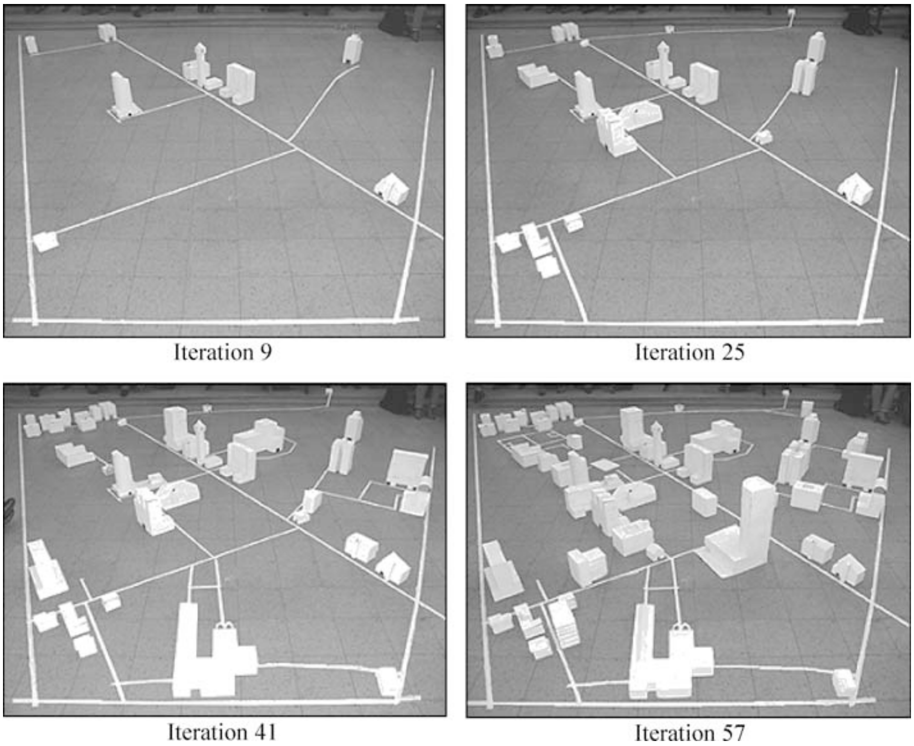


Self-Organization and the City, Figure 26

Another conceptualization of the public-collective SIRN sub-model. Each individual agent is subject to internal input – a cognitive map constructed by the mind/brain and external input –the legible information coming from the common reservoir, that is, the city of a planning team. The interaction between these two forms of input gives rise to a competition between alternative decision rules that ends up when one or a few decision rules "wins". The winning rule(s) is/are the order parameter(s) that enslave(s) the system. The emerging order parameter governs an external output, which in the case of a city is the agent's behavior and action in the city, and an internal output, which is an information feedback loop back to the agent's mind/brain

p. 15 in [53]) has consequently suggested two forms of information: *semantic information* which is information *with* meaning, versus *Shannonian information* which is "information with meaning exorcised". Haken further emphasizes that the process of self-organization implies "an enormous compression of information" (see pp. 25, 35, 151 in [52]).

Haken and Portugali [60] have studied information in the context of the city. They show that different elements of the city transmit different quantities of Shannonian information that can be practically measured by means of information *bits*, for example (Fig. 28). They further show that cognitive processes such as pattern recognition and categorization entail an enormous information compression thus affecting the quantity of the Shannonian information conveyed by the city (ibid) and that information compression is just one facet of the process – the other facet is *information inflation* [62]: In certain urban situations categorization might entail information compression while in others information inflation (Fig. 29). Infor-



Iteration 9

Iteration 25

Iteration 41

Iteration 57

Self-Organization and the City, Figure 27a
Four snapshots from a typical city game (at iterations 1, 15, 35, 57)

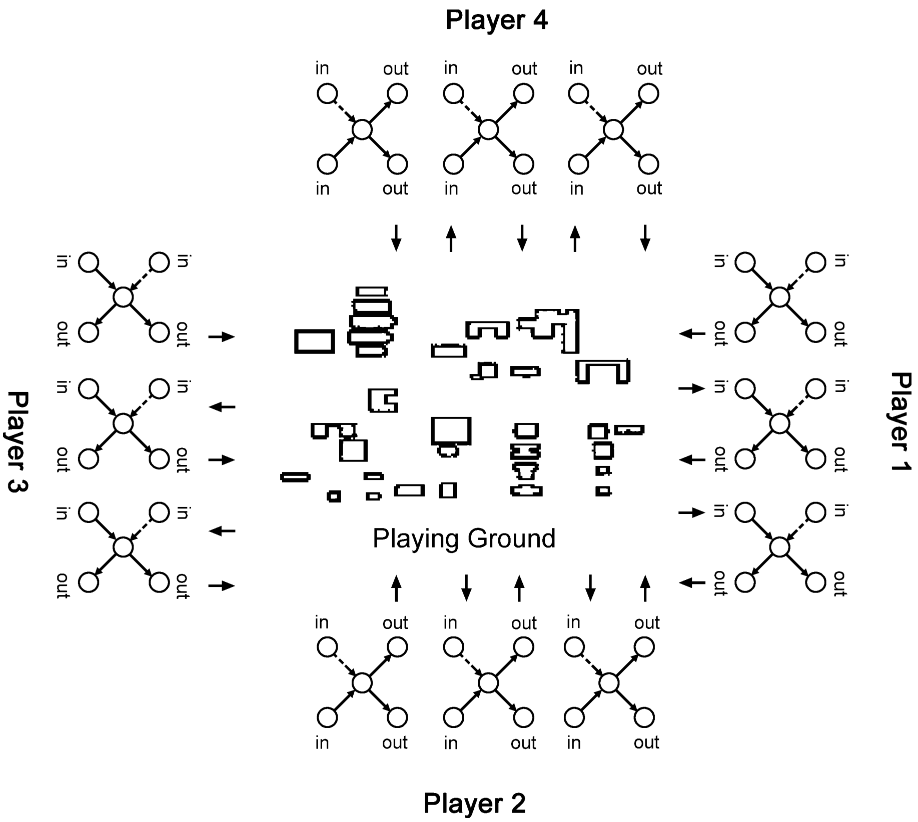
mation inflation and compression are thus two aspects of the process of *information adaptation* by which individuals and collectivities shape the city in a way that is adapted to the inhabitants' cognitive capabilities.

The notion information adaptation has far reaching implications to the above discussion about the scientific method and the science of cities: Self-organization as information compression implies a *complex* → *simple* model and thus an alternative Simon's *simple* → *complex* model. The process of information inflation, on the other hand, is in line with Simons' model. These two models are thus two facets of a single process of information adaptation that in some cases requires inflations while in others compression. Complexity theory shows that whatever are the opening conditions (complex or simple) a scientific approach is possible.

CogCity A central property of complex systems is the process of *circular causality* that typifies also the dynamics of cities: Thus in the SIRN model the interaction between the local/micro urban agents gives rise to the global structure of the city, which then feeds back and prescribes the behavior, interaction and action of the agents, and so on. Guided by Simon's *simple* → *complex* model, standard urban simulation models have become excellent tools to simulate the first part of this loop – the way local interactions give rise to a global structure – but they fail to describe the second, feedback part of the loop. CogCity (cognitive city) is a model that attempts to simulate the dynamic of cities as a process of circular causality [98].

CogCity is essentially a FACS model (above Sect. "AB and FACS cities") with several additions that make it an explicit SIRN, cognitive, urban simulation model. It differs from standard AB/CA urban simulation models in that the latter are essentially bottom-up in their structure (Fig. 30, *left*). CogCity, per contra, is characterized by an on-going interaction between top-down and bottom-up. Figure. 30, *right* describes a typical scenario: It starts top-down when an agent arrives to the city with a global cognitive map in mind; compares it to the global structure of the city and selects a local sub-area. Now starts the bottom-up process: the agent selects the empty cells in that local area; evaluates the appropriateness of the cells and their nearest neighbors in light of its needs and then takes a decision and action. In parallel, the properties of every cell are determined according to its relations to its nearest neighbors and so on.

In a regular AB/CA simulation the process ends here: the global outcome is recorded and mapped as the output



Self-Organization and the City, Figure 27b
A conceptualization of the city game

of this specific iteration and the model is ready for a new iteration. In a SIRN-CogCity model the process continues and feeds back to the global structure of the city that allows the top-down process in the next iteration: Firstly, the state of the various central places is determined. Secondly, peripheries are determined around central places. Thirdly, areas are defined or redefined. Fourthly, subareas are redefined. Fifthly, given areas and subareas, the global state of the city as a whole and its rank-size structure, is defined. The latter changes redefine the local membership state of each cell in the various infrastructure objects and become the externally represented input for a new agent in the next iteration, and so on in circular causality (Figs. 31, 32).

Planning

The link between self-organization and cities is contradictory. Firstly, since cities were always regarded as symbols of planned action – walls, roads, castles, fortresses, indicated a central authority that is capable of planning. Secondly, since planning as means to achieve a controlled order diametrically opposes self-organization as the spontaneous emergence of order. This is a seemingly contradiction, however, since cities are *dual* self-organizing systems with the implication that every urban agent is a planner at a certain scale. This view is supported by psychology and cognitive science that consider planning as one of the basic cognitive capabilities of humans [86].

Cognitive planning, that is, the ability to think, decide and act ahead, must be based on information about the future which by definition is partial and insufficient – a situation that according to Haken typifies also the process of pattern recognition as conceptualize by synergetics. Based on this analogy Haken [57] described decision situation in the context of planning as in Fig. 33.

This decision situation raises the question of 'How do people complement the unknown data?' According to Haken and Portugali, as in pattern recognition tasks here too, the unknown data is being supplied by means of associative memory [57,95], conceptual cognitive maps [98,99] and decision heuristics [95,127,128]. Table 1 specified several decision heuristics and their interpretation in the context of cities.

Self-Organization and the City, Table 1

Seven heuristics and their interpretation in the context of cities. For sources see [95]

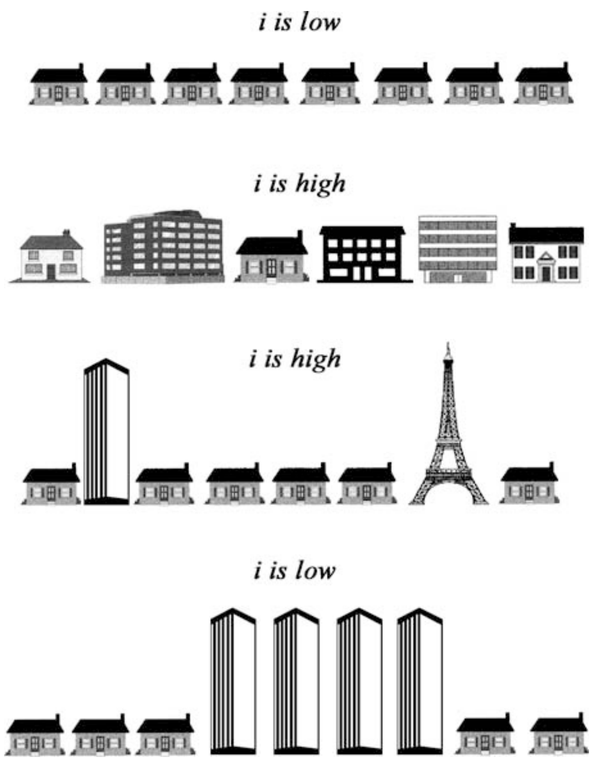
Heuristic	Description	Urban interpretation
Similarity	The similarity of two items is expressed as a function of their common and distinctive features.	The recognition of fundamental rules of urban composition, such as a grid layout, built form, or urban fabric, is performed through similarity.
Representativeness	The probability that an object or event belongs to a particular class is judged by the degree to which the description is representative of a stereotype.	Urban categories are identified on the basis of architectural stereotypes, such as church, skyscraper, boulevard, tower, park, arch, etc.
Availability	The probability of an event, or frequency of a class, is assessed by the ease with which instances or occurrences can be brought to mind, or recalled.	Availability would make universal symbols (such as Macdonald's signs, Stop signs, etc.) more easily identified and recalled.
Decision frame	The frame that a decision maker formulates of the problem (gain versus loss, etc.) is influenced by norms, habits, and personal characteristics of the decision maker.	Urban frames for decisions (congested versus free, public versus privet, etc.) depend on the cultural code of each agent (e. g. a tourist, a taxi driver, a policeman, etc.).
Anchoring	The tendency of people to make estimates by starting from an initial base value that is adjusted to yield the final answer.	City's internal representations can contain certain categories such as style, urban violence, town size, etc., which can be 'fired on' early in the process of cognition; once switched on, it stays on and is only eventually reprocessed.
Synergetic I: Collective effects	When facing complex decision situations people tend to rely on what other people are doing or saying.	Drivers, pedestrians, intra- and inter-urban immigrants, tend to 'follow the stream'. That is: to take decisions in line with what others are doing.
Synergetics II: Attention parameters effect	When facing complex decision situations people often employ several heuristics in a sequence. First, the attention parameter calls into use a heuristic. Then, when exhausted, another attention parameter heuristic emerges and so on.	Intra- and inter-urban immigrants, for example, often start with a given location decision heuristic (say synergetic I); if it doesn't work, they switch to an alternative heuristic and so on.

Haken and Portugali [95] have further suggested that complex processes of decision making in the context of city planning evolve according to their SIRN model. This suggestion was further elaborated by Portugali and Alfasi [105] who demonstrated empirically how this SIRN process practically takes place in the reality of *planning discourse* as it evolved among members of a planning team engaged in formulating urban policies concerning the development of the city of Beer Sheva, Israel (Fig. 34).

As a basic cognitive capability planning is intimately associated with the fact that humans are social creatures – people tend to plan together (e. g. families, friends, firms etc). Some planning decisions are thus made solitarily while others collectively. Planning is also a profession that is closely linked to the central authorities of society (municipal, regional, national governments etc.). We thus have three forms of planning – *solitary*, *collective* and *professional*.

The notion that cities are complex self-organizing systems thus implies a novel view on planning the essence of which is, first, that all three forms of planning (solitary, collective and professional) participate in the dynamic of cities. Second, that due to non-linearities that typify cities as complex systems, the act of a single solitary planner might affect the evolution of a city more than the planning act of a professional planning team. (For an example see [101]). Does that mean that due to self-organization there is no need for city planning? Not at all! – It means that we have to adopt a new perception of *plans as participants* in the overall urban dynamics. It also means that we have to adopt a new perception of urban dynamics as a complex interaction between many plans at different scales, or more specifically, between solitary, collective and professional, planning agents, each with its specific plan.

Can there be an administrative planning process that is built in line with the above? The answer suggested by Portugali and Alfasi is positive [105]: In a sequence of studies they have portrayed the principles of such self-planned city [95] and the way it can be applied to the reality of city planning law and structure of Israel [2]. Similarly to current planning systems it is a 3-layers system: the legislative, the judiciary and the executive. It differs in the following: First, its planning law refers to the qualitative relations between the various city objects (Fig. 35) and not to land use plans that assume to determine top-down the urban landscape. Second, it suggests a novel planning judicature



Self-Organization and the City, Figure 28
Different configurations and categorizations of buildings convey different quantities of Shannonian information. When all building are similar, information (*i*) is low. When they are different, *i is high* but difficult to memorize. When landmarks are added *i is high*, provided that they are located apart from each other; otherwise, *i is low*. Source: [60]

composed of spatially distributed "planning courts" conducted by professionals who have specialized in both law and planning. Their aim is to evaluate, accept or reject the plans proposed by all planning agents – solitary, collective or professional. Third, it suggests a separation of authorities that doesn't exist today in standard planning administrations. Fourth, it suggests a process of hermeneutic planning that enables phase transition and adaptation to new situations. Figure 36 describes the structure and operation of this self-organized planning system.

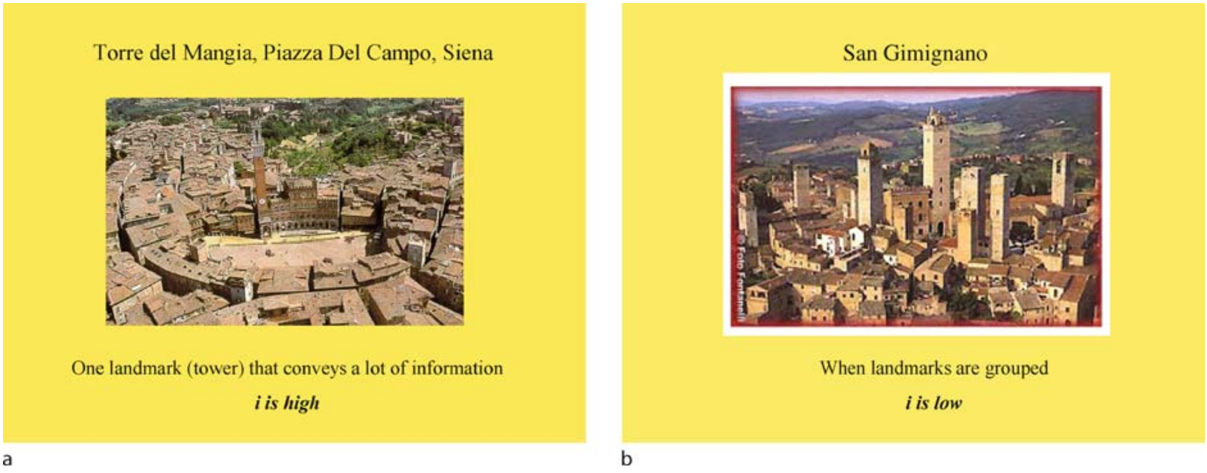
Prediction, Planning, Self-Organization and Cities

The first principle of the above planning system is that its planning laws refer to qualitative relations between the various city objects and not to land use plans. The reason is that land use plans are commonly based on predictions. This is problematic since prediction in the context of complex systems such as cities is associated with four fundamental properties. First, the nonlinearities that typify cities imply that one cannot establish predictive cause-effect relationships between some of the variables. Second, many of the triggers for change in complex systems have the nature of unpredictable mutations [4]; not because of lack of data, but because of their very nature. Third, unlike closed systems, in complex systems, the observer, with his/her predictions, is part of the system – a point made by Jantsch [71] more than two decades ago and largely ignored since then. In such a situation, predictions are essentially feed-forward loops, affecting the system and its future evolution with implications that include self-fulfilling and self-falsifying or self-defeating predictions [101].

From the above follows a dilemma: complex systems are in essence unpredictable and yet, the current practice of planning as well as planning administration and law are based on the ability to predict. In a recent paper (ibid) it was shown that this situation leads to planning paradoxes that are the result of phenomena of self-fulfilling and self-falsifying predictions. It was further shown that these phenomena are the result of the *feed-forward* and *feedback loops* that are typical of complex systems in general and of cities and regions in particular. The existence of such loops is one of the properties that make systems complex. Such loops are responsible to the situation by which a prediction or a plan, once produced, becomes a *participant* in the system's dynamic.

Another way to look at this issue is from the point of view of the distinction between Shannonian and semantic information (above, Sect "Information Compression, Inflation, Adaptation"): Predictions and plans are essentially kinds of information transmission. One can thus speak of Shannonian prediction and semantic prediction. In the first, the outcome of the prediction is independent of the receiver(s) while in the second it depends on the meaning attached to it by a receiver or receivers. A weather forecast is a good example for both: it has no effect on the climatic system, but it might affect the urban system – following the prediction people might behave in different ways that might entail phenomena of self-falsifying and self-fulfilling predictions as described above.

Planning theory has not as yet internalized the implications of complexity theory to city planning. For example, in the planning and decision support systems (PSS, DSS) that are currently discussed and built by proponents of the complexity paradigm, urban simulation models are assumed to function as sophisticated prediction devises [28,29,43]. The result is a discrepancy that to my mind characterizes the domain of urban and regional planning: On the one hand, planning theory, as well as the structure of planning law, practice and administration, are all based on the (usually implicit) assumption that


Torre del Mangia, Piazza Del Campo, Siena

One landmark (tower) that conveys a lot of information

i is high

a

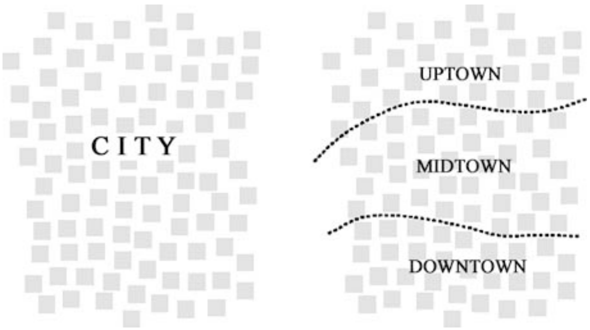
San Gimignano

When landmarks are grouped

i is low

b

Self-Organization and the City, Figure 28a, b
a An example of a good landmark. b An example of not very effective landmark



Self-Organization and the City, Figure 29
If all buildings in the city are different from each other, categorization entails information compression; if they are similar – information inflation

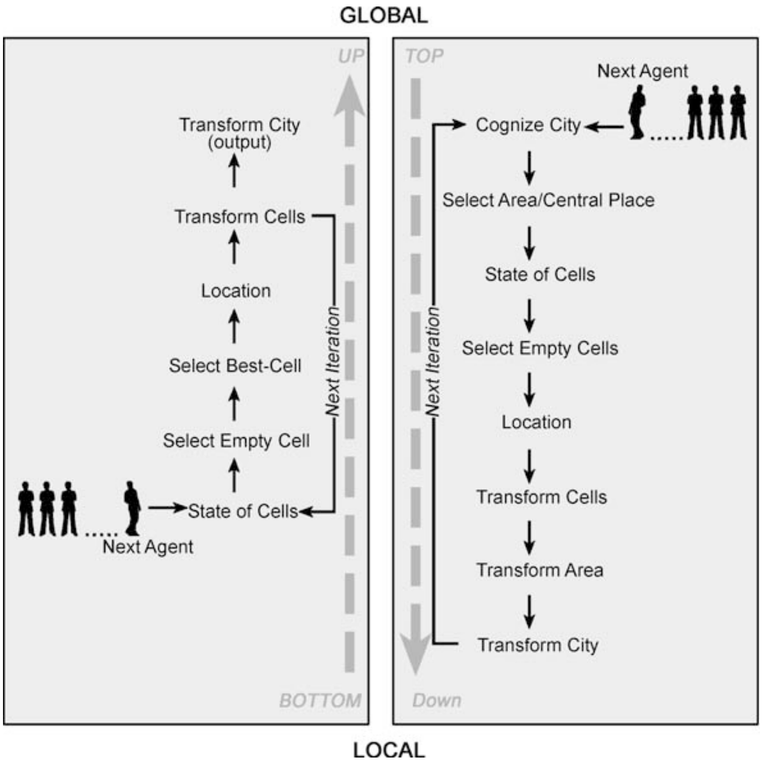
cities are essentially predictable entities; that given sufficient data, information and models, their future behavior is in essence predictable. On the other hand, current urban theory suggests that cities are complex, self-organizing and non-linear systems and that as a consequence their future behavior is in essence not predictable; even if sufficient information and data is collected and available [95].

Urbanism

In Sect. "Explicit Attempts to Define a City" we've defined the category "city" in terms of a family resemblance according to which a settlement becomes a "city" not by having some necessary and sufficient properties, but by being a member in a network of cities that has center, periphery etc. This view is in line with cognitive science's approach to concepts and categories. In the latter it is common also to distinguish between *basic level* categories and *super-ordinate* categories [116]. A 'chair', for instance, is a basic level category whereas 'furniture' a super-ordinate. The suggestion here is that a *city* is a basic level category whereas *urbanism* is a super-ordinate one, referring to the totality of cities ranging from their physical structure, architecture, economics, politics, social and cultural composition, and so on. The term 'urban revolution' (coined by Childe, [35]) thus implies a major transformation in society with the basic level category 'city' at its center.

Most complexity studies of cities have traditionally focused on specific aspects of cities – land-use, morphology, transportation, social segregation etc. – but not on the totality of city life which is what urbanism is all about. Why? Because they evolved mainly out of regional sciences' attempt to develop a scientific approach to cities and the consequent tendency to choose research issues that can be analyzed by reference to 'real world' data. The study of urbanism was thus left to the "soft" social theory approaches to cities[33,34,63,64].

This is rather unfortunate because 21st century world society is undergoing a major transformation with urbanism at its center: Massive rural-urban migration and demographic processes entailed a situation by which cities such as Mexico City, Bombay (Mumbai) and Sao Paolo grew from 8.8 million, 6.2 million and 8.3 million respectively in 1970 to over 20 million, over 16 million and again over 18 million today; for the first time in human history, the number of people living in cities is crossing 50% of the world's population and the process is still on. In the last



Self-Organization and the City, Figure 30
A cognitive (*right*) vs non-cognitive (*left*) AB/CA urban simulation model. Source: [98]

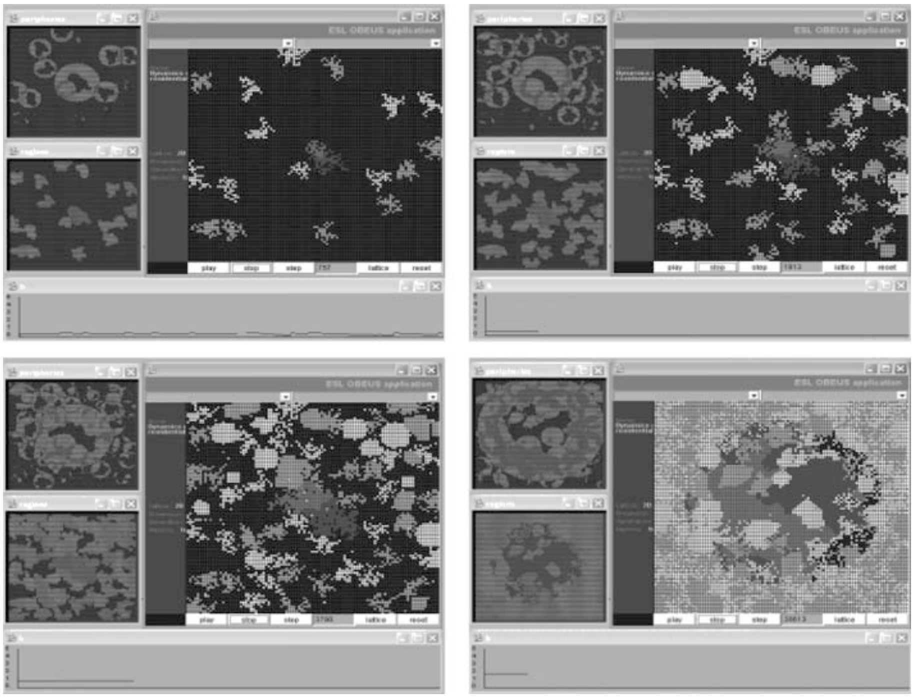
few decades we've witnessed the emergence of *world cities*, or *global cities*, that form the centers for the globalization process.

These quantitative processes are associated with several qualitative processes: a process of privatization that leads to the decline of the welfare nation-state; the emergence of a civil society that takes over many of the past duties and functions of the nationalist welfare state; the crucial problems of many (post)modern counties are no longer classical national problems (e. g. national self-determination, national boundaries etc.), but rather the problems of cities. The events of September 11 and the ensuing wars in Afganistan and the Middle East are tragic indications to the *urbanization of war*. Finally, the process of globalization is making some world cities more dominant than the states within which they exist thus repressing the nation states.

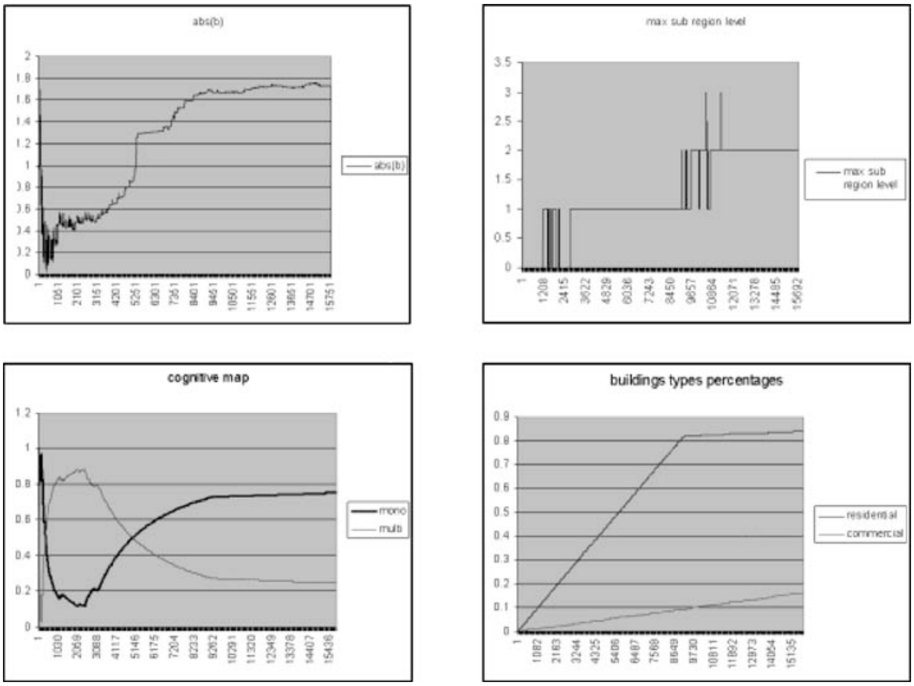
All the above indicates the more fundamental change: According to Lefebvre [79], its essence is that urbanism is replacing industrialization as the dominant force in society. My view is that the essence of this change is that urbanism is challenging nationalism as the order parameter of modern society [101].

Complexity and Urbanism Complexity studies of cities, with their focus on the short-term dynamics of cities and of national systems of cities are indeed highly advanced in terms of mathematical formalism and data analysis but rather anachronistic in terms of the issues studied; as such they have so far said very little on the dramatic urban phenomena of 21st century. Can they say more about the issue of urbanism? The answer is yes! And for several reasons: To my mind the "deeper messages" of complexity theories is that they have discovered properties in matter hitherto assigned to life, art and society [91]. It is not surprising therefore that complexity theories, particularly synergetics, bear many similarities to social theory and philosophy and, as a consequence, several of the notions that originated in the study of complex systems can be related to similar notions that originated in the domain of social theory [100]:

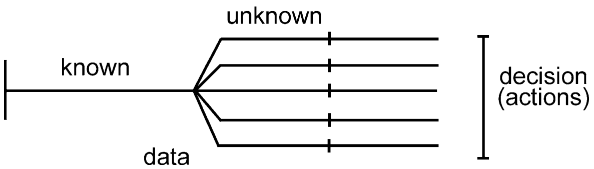
- Both are essentially systemic and even holistic.
- Both tend to conceptualize 'development' and 'evolution' in terms of abrupt changes rather then a smooth progression. In social theory the common terms for an abrupt change is (social/political/cultural) 'revolution',



Self-Organization and the City, Figure 31
Preliminary results from an evolving scenario simulated by CogCity: The central screen in each of the four snapshots shows the evolving spatial distribution of various kinds of agents, the top left screen the evolution of centers and sub-centers, while the bottom left, the evolving cognitive maps of agents. Source: [98]



Self-Organization and the City, Figure 32
Preliminary results from an evolving scenario simulated by CogCity – the graphs. Source: [98]



Self-Organization and the City, Figure 33
Decisions in the city are characterized by insufficient data. In such a reality the known data may be complemented in a variety of ways. Each of theses ways might entail a different decision and action. Source: [57,95]

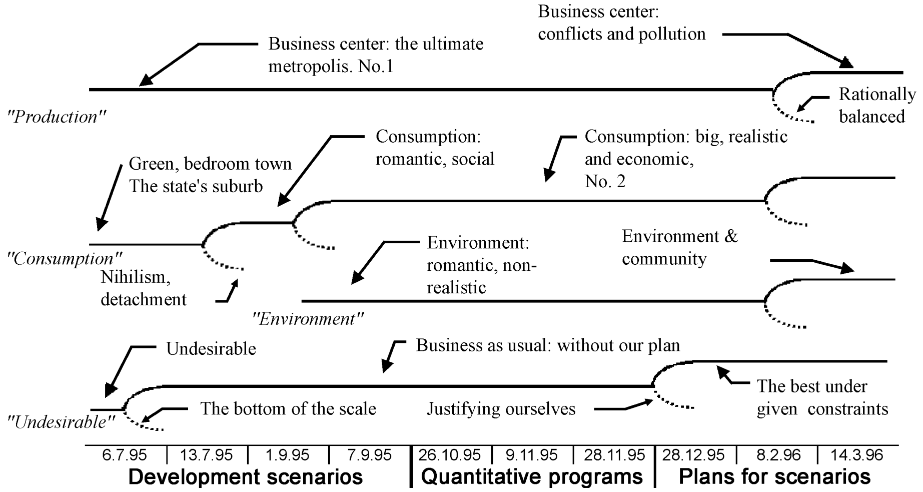
while in the language of complexity 'bifurcations' and 'phase transitions' (that reminds one of Gould's and Eldredge's, [46] *punctuated equilibrium*).

- Synergetics' notion of 'order parameter' is similar to social theory's notion of 'mode of production'.
- Synergetics' notions of 'enslavement' and 'circular causality' are close to social theory's notions of 'social reproduction' and 'socio-spatial reproduction' [44,80].
- Complexity's view of systems in 'a far from equilibrium condition' comes close to postmodernism's recent emphasis on viewing reality as ever changing and transforming; hence the general popularity of notions such as 'chaos' and 'butterfly effect'.

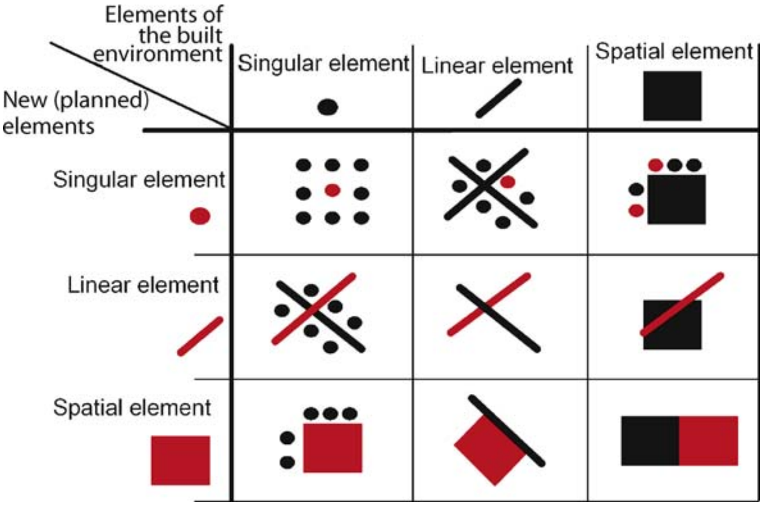
Several writers have already responded to these similarities from the perspective of the sciences, philosophy, media/cultural critics and modern and postmodern social theory [37,75,76,100,113]. A preliminary attempt has also been made to employ synergetics as a complexity theory of urbanism [95,101]. That is, to interpret the current changes in cities and urbanism in terms of synergetics along the following scenario: the combined force of rapid population growth, urban expansion and technological change throughout the 20th century acted as a control parameter. Toward the end of the 20th century and at the beginning of the 21st, we are witnessing a bifurcation and phase transition followed by a competition between the newly emerging *urban order parameter* and the old nationalist one. My personal view is that what we see emerging today out of this competition is not the replacement of nationalism as an order parameter by urbanism as an order parameter, but the *urbanization of nationalism*.

Future Directions

Looking in retrospect at more than two and a half decades of complexity theory studies of cities, one can now appreciate some of its major achievements: First, the link between cities and complexity theory gave urban studies a strong theoretical basis it never had before. The fact that complexity theory was applied to a large number of domains gave urban studies a wide context and many sources of inspiration. The fact that complexity theory comes with a rich and strong mathematical formalism gave urban studies a sound methodological background. The attempt

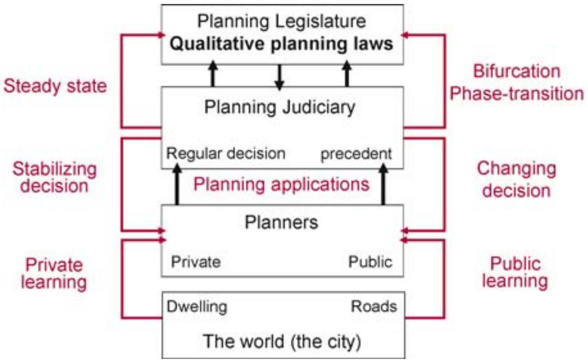

Business center: conflicts and pollution
Business center: the ultimate metropolis. No.1
"Production"
Rationally balanced
Consumption: romantic, social
Consumption: big, realistic and economic, No. 2
Green, bedroom town The state's suburb
"Consumption"
Nihilism, detachment
Environment: romantic, non-realistic
Environment & community
"Environment"
Undesirable
Business as usual: without our plan
The best under given constraints
"Undesirable"
The bottom of the scale
Justifying ourselves
6.7.95
13.7.95
1.9.95
7.9.95
26.10.95
9.11.95
28.11.95
28.12.95
8.2.96
14.3.96
Development scenarios
Quantitative programs
Plans for scenarios

Self-Organization and the City, Figure 34
Bifurcation diagram of the planning discourse. Each alternative is represented with a continuous line along the time axis (the X-axis). A horizontal line represents an order state during which the alternative scenario maintains a certain image and possesses certain attributes. Bifurcation points indicate a shift from one order state to another. The broken lines represent optional order states that were not actualized. Source: [105]



Self-Organization and the City, Figure 35

Proposition for a new planning structure for a self-planned city: The interrelations between urban elements provide the basis for planning law. A singular urban element might be a building, a linear one might be a road, while an example for a spatial urban element is a park



Self-Organization and the City, Figure 36

Proposition for a new planning structure for a self-planned city: The system is built of private planners (the inhabitants of the city) and professional planners. Each of them might submit a plan to the planning judiciary. In the latter the "planning judge" takes decisions according to the planning law as determined by the planning legislature. Unlike the current structure, there is a clear separation of authorities

to transform the study of cities into a science of cities is today closer than ever.

Complexity theory has given us a new insight to our understanding of the dynamics of cities. According to Batty [16] the most important contribution is that complexity studies of cities have verified the intuitive views of Jean Jacobs [70] and Alexander [1], namely, that the complex entity "city", with its variety of different land-uses, socio-spatially and culturally segregated communities, transportation networks and all the rest, is an outcome of "bottom-up" processes: The local interaction between agents at local scale, conducted by very few and simple rules gives rise the complexity we term 'city'.

At the same time, however, it must be admitted that the potential contribution of complexity theories to urbanism, planning and urban design has yet to be realized. The complexity approach has indeed given the bottom-up views on the nature of cities a strong mathematical formalism that can be quantified by real data. But this focus on the local, the bottom-up and the quantifiable was not without price: Cities and urbanism of the 21st century are in the midst of a dramatic transformation, new forms of cities are emerging – world cities, global cities, megacities, and yet the vast majority of complexity studies still focus on the old traditional quantitative urban questions, leaving the qualitative grand urban issues to the "non-scientific" social theory oriented urban studies. Can complexity theories of cities contribute? The answer suggested above is yes! SIRN is one approach in this direction and the field is ripe for others.

The same applies to planning. In the literature on planning theory it is common to make a distinction between *planning theory* versus *theory in planning* [40]. That is, between a theories about how to plan and theories about urban and regional dynamics that planners can use during the planning process. Examining complexity theories of cities from this perspective we see that they are very innovative with respect to theory in planning, but very conservative when it comes to theory of planning: The vast majority of studies simply ignore the implications of complexity to urban, regional and environmental planning.

Why do we need a complexity theory of planning? The answer is twofold: First, standard planning theory was developed in the 1950s and 1960s hand in hand with what we consider today as anachronistic urban theory: Both are based on the (usually implicit) assumption that cities are in essence simple, mechanistic systems that given sufficient data and advanced technologies, their future behavior is predictable and hence controllable. As we've seen above, complexity theories tell us a different story: Cities are complex self-organized systems that are in essence unpredictable and controllable; even if sufficient data and the most advanced technologies are at hand. From here follow a whole set of new and interesting questions: what is the role of planning in a complex system? Are all parts and components of the system unpredictable? In the above we've suggested some preliminary answers – but new ones must still come.

Finally it is important to mention the issue of extreme events in cities. The rapid processes of urbanization cities underwent in the last few decades and the 'urbanization of war' made cities rather vulnerable areas in cases of extreme events. The question of how cities and their inhabitants behave and respond to extreme events, is a pressing social issue that already started to capture the attention of students of complexity theory of cities and urbanism.

Bibliography

1. Alexander C (1965) A city is not a tree. Architectural Forum April-May, 58–62
2. Alfasi N, Portugali J (2007) Planning rules for a self-planned city. Planning Theory 6(2):164–182
3. Allen PM, Strathern M (2004) Complexity: The integrated framework for integrated models of urban and regional systems. A talk delivered at a conference on The Dynamics of Complex Urban Systems, November 4–6, 2004, Monte Verita
4. Allen PM (1997) Cities and Regions As Self-Organizing Systems: Model of Complexity. Routledge, London
5. Allen PA (1981) The evolutionary paradigm of dissipative structures. In: Jantsch E (ed) The Evolutionary Vision. Westview Press, Boulder, pp 25–71
6. Allen P, Sanglier M (1981) Urban evolution, self-organization and decision making. Environ Planning A 13:169–183
7. Allen P, Sanglier M, Engelen G, Boon F (1985) Towards a new synthesis in the modeling of evolving complex systems. Environ Planning B Planning Des 12:65–84
8. Alonso W (1965) Location and Land Use. Harvard University Press, Cambridge, MA
9. Andersson C, Hellervik A, Lindgren K, Hagson A, Tornberg J (2003) Urban economy as a scale-free network. Phys Rev E 68(3):036124
10. Auerbach F (1913) Das Gesetz der Bevolkerungskoncentration. Petermanns Geogr Mitt 59:74–76
11. Bak P, Chen K, Creutz M (1989) Self-organized criticality in the game of life. Nature 342:780–782
12. Bak P, Chen K (1991) Self-organized criticality. Sci Am 28: 26–33
13. Barabási A-L (2002) Linked: The new science of networks. Perseus, Cambridge
14. Barabási A-L, Réka A (1999) Emergence of scaling in random networks. Science 286:509–512
15. Bartlett FC (1932/1961) Remembering: A Study in Experimental and Social Psychology. Cambridge University Press, Cambridge
16. Batty M (2005) Cities and Complexity: Understanding Cities with Cellular Automata, Agent Based Models and Fractals. MIT Press, Cambridge, Mass
17. Batty M (1997) Cellular automata and urban form: a primer. J Am Planning Assoc 63(2):266–274
18. Batty M (1996) Urban Evolution on the Desktop: Simulations Using Extended Cellular Automata. Unpublished paper, Centre for Academic Spatial Analysis, UCL
19. Batty M, Longley P (1994) Fractal Cities. Academic Press, London
20. Batty M, Xie Y (1999) Self-organized criticality and urban development. In: Portugali J (ed) Population, Environment and Society on the Verge of the 21st Century, pp 109–124
21. Batty M, Xie Y (1994) From cells to cities. Environ Planning B: Planning Des 21:531–548
22. Benenson I, Torrens PM (2004) Geosimulation: Automata based modeling for urban phenomena. Wiley, London
23. Benguigui L (1995) A fractal analysis of the public transportation system of Paris. Environ Planning A 27:1147–1161
24. Benguigui L, Czamanski D, Marinov M, Portugali J (2000) When and where is a city fractal. Environ Planning B 27(4): 507–519
25. Blanchard P, Volchenkov D (2007) Scale-free Segregation in Transport Networks. arXiv:0710.1592
26. Braudel F (1993) A History of Civilizations, translated by Richard Mayne. Penguin Books, New York
27. Blumenfeld-Lieberthal E (2005) Dynamics of Urban Morphology in the Tel-Aviv Metropolitan Area. Ph.D thesis, Technion, Haifa
28. Brail RK (2006) Planning support systems evolving: When the rubber hits the road. In: Portugali J (ed) Complex Artificial Environments, Complexity series. Springer, Heidelberg, pp 307–317
29. Brail RK, Klosterman RE (eds) (2001) Planning Support Systems. ESRI Press, New York
30. Burgess B (1925) The growth of the city: An introduction to a research project. In: Park R, Burgess EW, McKenzie RD (eds) The City. Chicago University Press, Chicago, pp. 47–62
31. Burgess EW (1926/1968) The Urban Community. Greenwood Press, New York
32. Burgess EW (1927) The determination of gradients in the growth of the city. American Sociological Society Publications 21:178–184
33. Castells M (1977) The Urban Question. MIT Press, Cambridge, MA
34. Castells M (1996) The Rise of the Network Society. Blackwell Publishers, Malden, MA
35. Childe VG (1950) The urban revolution. Town Planning Rev 21:3–17

36. Christaller W (1933/1966) Central Places in Southern Germany. Prentice Hall, Englewood Cliffs, NJ
37. Cilliers P (1998) Complexity and Postmodernism. Routledge, London
38. Dalton N, Peponis J, Conroy-Dalton R (2003) To tame a TIGER one has to know its nature: extending weighted angular integration analysis to the description of GIS road-centerline data for large scale urban analysis. In: Hanson J (ed) Proceedings of the 4th International Space Syntax Symposium. University College London, London, pp 65.1–65.10
39. Dendrinos DS, Sonis M (1990) Chaos and Socio–Spatial Dynamics. Springer, New York, Berlin
40. Faludi A (1973) A Reader in Planning Theory. Pergamon Press, Oxford
41. Figueiredo L, Amorim L (2005) Continuity lines in the axial system. In: Van Nes A (ed) 5th International Space Syntax Symposium. TU Delft, Faculty of Architecture, Section of Urban Renewal and Management, Delft, pp 161–174
42. Garling T, Golledge RG (eds) (1993) Behavior and Environment. North-Holland, Amsterdam, London, New York, Tokyo
43. Geertman S, Stillwell J (eds) (2003) Planning Support Systems in Practice. Springer, Heidelberg
44. Giddens A (1984) The Constitution of Society: Outline of the Theory of Structuration. University of California Press, Berkeley
45. Golledge RG, Timmermans H (eds) (1988) Behavioral Modeling in Geography and Planning. Croom Helm, London, New York, Sydney
46. Gould SJ (1980) The Panda's Thumb. Norton, New York
47. Gregory D (1994) Geographical Imaginations. Blackwell, Cambridge, MA
48. Haag G, Muntz M, Pumain D, Saint-Julien T, Sanders L (1992) Interurban migration and the dynamics of a system of cities. Environment and Planning A 24:181–198
49. Haken H (1979) Pattern formation and pattern recognition – an attempt at a synthesis. In: Haken H (ed) Pattern Formation by Dynamical Systems and Pattern Recognition. Springer, Berlin, pp 2–13
50. Haken H (1983) Synergetics, an Introduction, 3rd edn. Springer, Berlin
51. Haken H (1987) Advanced Synergetics: An Introduction (2nd print). Springer, Berlin, New York
52. Haken H (1988/2000) Information and Self-Organization: A Macroscopic Approach to Complex Systems. Springer, Berlin
53. Haken H (1990) Synergetics of Cognition. Springer, Berlin
54. Haken H (1991/2004) Synergetic Computers and Cognition. Springer, Berlin
55. Haken H (1993) Synergetics as a strategy to cope with complex systems. In: Haken H, Mikhailov A (eds) Interdisciplinary Approaches to Non-Linear Complex Systems. Springer, Berlin
56. Haken H (1996) Principles of Brain Functioning: A Synergetic Approach to Brain Activity, Behavior and Cognition. Springer, Berlin
57. Haken H (1998) Decision making and optimization in regional planning. In: Beckmann MJ, Johansson B, Snickars F, Thord R (eds) Knowledge and Networks in a Dynamic Economy. Springer, Berlin
58. Haken H, Portugali J (1995) A synergetic approach to the self-organization of cities. Environ Planning B: Planning Des 22:35–46
59. Haken H, Portugali J (1996) Synergetics, inter-representation networks and cognitive maps. In: Portugali J (ed) The Construction of Cognitive Maps. Kluwer, Dordrecht, pp 45–67
60. Haken H, Portugali J (2003) The face of the city is its information. Environ Psychol 23:385–408
61. Haken H, Portugali J (in preparation) Synergetics and captivity in a city
62. Haken H, Portugali J (in preparation) Information adaptation
63. Harvey D (1985) The Urbanization of Capital. Basil Blackwell, Oxford
64. Harvey D (1996) Justice, Nature and Geography of differences. Basil Blackwell, Oxford
65. Hillier B (1999) Space is the Machine: A Configurational Theory of Architecture. Cambridge University Press, Cambridge
66. Hillier B, Hanson J (1984) The social logic of space. Cambridge University Press, Cambridge
67. Hillier B, Iida S (2005) Network effects and psychological effects: a theory of urban movement. In: Cohn A, Mark D (eds) Spatial Information Theory. Lecture Notes in Computer Science, vol 3603. Springer, Berlin, pp 473–490
68. Hu MB, Wang WX, Jiang R, Wu QS, Wang BH, Wu YH, (2006) Urban Traffic Dynamics: A Scale-Free Network Perspective, http://arxiv.org/abs/physics/0606086v1
69. Isard W (1956) Location and Space Economy. Published jointly by the Technology Press of Massachusetts Institute of Technology and Willey, New York
70. Jacobs J (1961) The Death and Life of Great American Cities. Penguin Books, England
71. Jantsch E (ed) (1981) The Evolutionary Vision. Westview Press, Boulder, CO
72. Jiang B (2007) A topological pattern of urban street networks: Universality and peculiarity. Physica A 384:647–655
73. Jian B (2006) Small world modeling for complex geographic environments. In: Portugali J (ed) Complex Artificial Environments: simulation, cognition and VR in the study and planning of cities. Springer, Berlin
74. Johnson M (1987) The body in the mind: The bodily basis of meaning, imagination, and reason. The University of Chicago Press, Chicago
75. Johnson (2001) media/cultural critics – postmodern complexity
76. Kellert SH (1993) In the Wake of Chaos. University of Chicago Press, Chicago
77. Kitchin R, Freundschuh S (2000) Cognitive Mapping: Past, Present and Future. Outledge, London
78. Lakoff G (1987) Women, Fire and Dangerous Things: What Categories Reveal About the Mind. The University of Chicago Press, Chicago, London
79. Lefebvre H (1970) La Révolution Urbaine. Gallimard, Paris
80. Lefebvre H (1974) The Production of Space. English translation, 1995. Blackwell, Oxford
81. Lösch A (1954) The Economics of Location. Yale University Press, New Haven
82. Lynch K (1960) The Image of the City. MIT Press, Cambridge, MA
83. Mandelbrot BB (1983) The Fractal Geometry of Nature. Freeman, San Francisco
84. Milgram S (1967) The Small World Problem. Psychol Today May:60
85. Miller GA (1956) The magic number seven, plus or minus two:

Some limits on our capacity for processing information. Psychol Rev 63(2):81–97

86. Morris R, Ward G (eds) (2005) The cognitive psychology of planning. Psychology Press, Hove, East Sussex, New York
87. Nicolis G, Prigogine I (1977) Self-Organization in Nonequilibrium Systems: From Dissipative Structures to Order Through Fluctuations. Wiley-Interscience, New York
88. Park RE (1925/1967) The City. Chicago University Press, Chicago
89. Porta S, Crucitti P, Latora V (2005) The Network Analysis of Urban Streets: A Primal Approach, http://arxiv.org/abs/physics/0506009
90. Popper KR (1959) The Logic of Scientific Discovery. Hutchinson, London
91. Portugali J (1985) Parallel currents in the natural and social Sciences. In: Portugali J (ed) Links Between Natural and Social Sciences. A special theme issue. Geoforum 16(2), pp 227–238
92. Portugali J (1990) Preliminary notes on social synergetics, cognitive maps and environmental recognition. In: Haken H, Stadler M (eds) Synergetics of Cognition. Springer, Berlin, pp 379–392
93. Portugali J (1994) Theoretical speculations on the transition from nomadism to monarchy. In: Finkelstein I, Na'aman N (eds) From Nomadism to Monarchy. Yad Izhak Ben-zvi, Jerusalem, pp 203–217
94. Portugali J (1996) Inter-representation networks and cognitive maps. In: Portugali J (ed) The Construction of Cognitive Maps. Kluwer, Dordrecht, Boston, London, pp 11–43
95. Portugali J (2000) Self-Organization and the City. Springer, Heidelberg
96. Portugali J (2002) The Seven Basic Propositions of SIRN (Synergetic Inter-Representation Networks). Nonlinear Phenom Complex Syst 5(4):428–444
97. Portugali J (2003) SIRN (Synergetic Inter-Representation Networks), Artifacts and Snow's two cultures. In: Tschacher W, Dauwalder J-P (eds) Dynamical System Approaches to Embodied Cognition. World Scientific, Singapore, pp 277–294
98. Portugali J (2004) Toward a cognitive approach to urban dynamics. Environ Planning B: Planning Des 31:589–613
99. Portugali J (2005) Cognitive Maps Are over 60. In: Cohn AG, Mark DM (eds) COSIT 2005. LNCS, vol 3693. Springer, Berlin, pp 251–264
100. Portugali J (2006) Complexity theory as a link between space and place. Environ Planning A 38:647–664
101. Portugali J (2006) The scope of complex artificial environments. In: Portugali J (ed) Complex Artificial Environments: Simulation, cognition and VR in the study and planning of cities. Springer, Heidelberg
102. Portugali J (ed) (1992) Geography, Environment and Cognition. A special theme issue. Geoforum 23(2)
103. Portugali J (ed) (1996) The Construction of Cognitive Maps. Kluwer, Dordrecht
104. Portugali J (ed) (2006) Complex Artificial Environments: Simulation, cognition and VR in the study and planning of cities. Springer, Heidelberg
105. Portugali J, Alfasi N (2008) An approach to planning discourse analysis. Urban Studies, Forthcoming
106. Portugali J, Benenson I, Omer I (1994) Socio–spatial residential dynamics: stability and instability within a self-organizing city. Geograph Anal 26(4):321–340
107. Portugali J, Benenson I, Omer I (1997) Spatial cognitive dissonance and sociospatial emergence in a self-organizing city. Environ Planning B: Planning Des 27:263–285
108. Portugali J, Haken H (1992) Synergetics and cognitive maps. In: Portugali J (ed) Geography, Environment and Cognition. A special issue. Geoforum 23(2), pp 111–130
109. Prigogine I (1980) From Being to Becoming. Freeman & Co, San Francisco, CA
110. Prigogine I, Stengers I (1984) Order Out of Chaos. Bantam, New York
111. Pumain D (ed) (2005) Hierarchy in Natural and Social Sciences, Methodos Series (Hardcover). Springer, Heidelberg
112. Pumain D, Saint-Julien T, Sanders L (1987) Application of a dynamic urban model. Geograph Anal 19(2):152–166
113. Rasch W, Wolfe C (2000) Observing Complexity. University of Mennesota Press, Minneapolis
114. Ratti C (2004) Space syntax: some inconsistencies. Environ Planning B Planning Des 31:487–499
115. Rosch E (1999) Reclaiming concepts. J Consciousness Stud 6(11–12):61–77
116. Rosch E, Mervis C, Gray W, Johnson D, Boyes-Braem P (1976) Basic objects in natural categories. Cogn Psychol 8:382–439
117. Salingaros N (2006) A Theory of Architecture. Umbau-Verlag, Solingen, Germany
118. Salingaros N (2005) Principles of Urban Structure. Techne Press, Amsterdam, Holland
119. Sanders L (1992) Systems de Villes et Synergetique. Anthropos, Paris
120. Sanders L, Pumain D, Mathian H, Guerin-Pace F, Bura S (1997) SIMPOP: a multiagent system for the study of urbanism. Environ Planning B Planning Des 24:287–306
121. Sanglier M, Allen P (1989) Evolutionary models of urban systems: an application to the Belgian provinces. Environ Planning A 21:477–498
122. Schelling T (1974) On the ecology of micro-motives. In: Marris R (ed) The Corporate Society. Macmillan, London
123. Shannon CE, Weaver W (1959/1963) The Mathematical Theory of Communication. University of Illinois Perss, Illinois
124. Simon H (1979) The Science of the Artificial. MIT Press, Cambridge, MA
125. Snow CP (1964) The Two Cultures and a Second Look. Cambridge University Press, Cambridge
126. Stadler M, Kruse P (1990) The self organization perspective in cognition research: historical remarks and new experimental approaches. In: Haken H, Stadler M (eds) Synergetics of Cognition. Springer, Berlin, pp 32–52
127. Tversky A, Kahneman D (1974) Judgement under uncertainty: Heuristics and biases. Science 185:1124–1131
128. Tversky A, Kahneman D (1981) The framing of decision and psychology of choice. Science 211:4538
129. von Thünen JH (1826/1966) von Thünen's Isolated State. An English translation (Hall P (ed)) Pergamon, Oxford
130. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393:440–442
131. Weidlich W, Haag G (1983) Concepts and Models of a Quantitative Sociology. Springer, Berlin, New York
132. Weidlich W (1987) Synergetics and social science. In: Graham R, Wunderlin A (eds) Lasers and Synergetics. Springer, Berlin, pp 238–256
133. Weidlich W (1994) Synergetic modeling concepts for sociodynamics with application to collective political opinion formation. J Math Sociol 1(4):267–291

134. Weidlich W (1999) From fast to slow processes in the evolution of urban and regional settlement structures: the role of population pressure. In: Portugali J (ed) Population, Environment and Society on the Verge of the 21st Century. A special theme issue, Discrete Dynamics in Nature and Society, pp 137–147
135. Wirth L (1938) Urbanism as a Way of Life. Am J Sociol 64: 1–24
136. Wittgenstein L (1953) Philosophical Investigations. Blackwell, Oxford. Translated by G. E. M. Anscombe
137. Wu J, Gao Z, Sun H, Huang H (2004) Urban Transit System as a Scale-Free Network. Mod Phys Lett B 18(19–20):1043–1049
138. Zipf GK (1949) Human Behaviour and the Principle of Least-Effort. Addison-Wesley, Cambridge

Self-Organization in Clinical Psychology

GÜNTER SCHIEPEK[1], VOLKER PERLITZ[2]
[1] Institute of Synergetics and Psychotherapy Research, Paracelsus Medical University, Salzburg, Austria
[2] Klinik für Psychosomatik, Universitätsklinikum der RWTH Aachen, Aachen, Germany

Article Outline

Definition of the Subject

Clinical psychology is a sub-discipline of psychology engaged in the description, classification, explanation, and treatment of mental disorders. The primary focus is on psychological methods, models, and topics such as behavior, cognition, emotion, and social interaction with substantial overlap with related areas in psychiatry, psychosomatics, or behavioral medicine. Yet, the main stream in clinical psychology views the etiology of mental disorders, their time courses and susceptibility to psychological treatment still through the magnification glass of linear input-output philosophy of human functions. Owing to this paradigm, linear combinations of variables (as inner conflicts, irrational cognitions, or stressors) trigger the development of psychiatric diseases or disorders in genetically predisposed individuals. Therefore, linear multivariate regression models are assumed be able to predict the probability of falling ill or suffering from disorders. As an important field in clinical psychology, psychotherapy research defines randomized controlled trials as the golden standard of outcome research. Here, patients randomly assigned to different treatment modalities are being compared with respect to the outcome of different tests. In this regard, the input (treatment) is thought to determine the outcome (treatment effects).

Contrary, or rather, supplementing this line of research is the scientific paradigm of self-organization, i. e. the functioning of complex nonlinear systems with circular causality at its center. Gestalt psychology, traditionally concerned with patterns ("Gestalts") in perception, human behavior and interaction (e. g., those prevalent in group dynamics, Lewin [25]) focuses on such self-organization processes. Gestalt psychologists like Wolfgang Köhler (e. g. [28]), Wolfgang Metzger, Max Wertheimer, Kurt Lewin and others can be seen as direct predecessors of modern complexity researchers in psychology [67]. Another root of this development is Jean Piaget's equilibration theory of action-cognition patterns (schemata) describing assimilation-accommodation-cycles of these schemata [49]. During these processes, input from the inner and outer environment assumes the role of disturbing stimulation of individual system dynamics. A third important line of thinking in circular causality comes from anthropological medicine. The "Gestaltkreis" integrates feedback loops between sensorial and actional systems on the one side, and individual and environmental systems on the other side (ecosystemic approach) [84].

Introduction

During the past decades in clinical psychology, it was particularly the transdisciplinary approach of *synergetics* [16] which inspired a specific nonlinear and complexity research on cognition [17,79]), social interaction [41,77], etiology and dynamics of mental diseases (e. g., [57,78]), and psychotherapy (for an overview see Haken and Schiepek [21]). Synergetics describes, measures, and explains the autonomous processes of pattern formation and pattern transitions in complex nonlinear systems. Founded on Haken's fundamental discovery that these processes do not depend of the matter of the systems they occur in, synergetics became one of the most important inspirations to many scientific fields and topics. Especially, Haken early transferred synergetics to brain research (e. g., [3]), since the brain is an outstanding example of a complex, self-organizing system. Today, it is widely accepted that the brain and a serial computer not only differ profoundly, but there is almost nothing they share: No wonder in light of the more than 10^{11} nonlinear interconnected neurons forming a dynamic mega-network of neural networks with essential features like arrays of emerging and submerging synchronizations, its flexibility and ever changing pattern formation, working at the edge of chaos, or realizing combined (activating and inhibiting) feedback mechanisms following the principles set forth by synergetics which describes the laws of self-organizing systems [18,19].

Taking a closer look at most of the phenomena clinical psychology is concerned with it becomes obvious that they appear to be of dynamic nature. Human development processes, human change and learning processes, the dynamics and prognosis of mental disorders, problems manifesting in social systems like couples, families, teams, or the question of how psychotherapy works: Self-organization is a ubiquitous entity.

Dynamic Diseases

Mental disorders are characterized by specific dynamic patterns, mirroring "endogenous" and common features of a disorder (like the repetitive phases of unipolar major depression or the bipolar phases of bipolar disorders, oscillating between mania and depression), as well as the effects of an individual life-style including individual coping and treatment efforts. Mental disorders can be conceived as highly structured and coherent states which enslave and thus impair the individual's mental and social functioning. Following the "enslaving principle", emerging order parameters reduce the degrees of freedom in the behavior of the single parts of a system. There is phenomenal evidence that this is the case in many mental disorders. Obsessive-compulsive disorder patients coerced to repeat unwished thoughts or rituals are just but one most impressive example. On the brain level, such pathological states correspond with abnormal synchronization in specific neural networks impairing brain functions. In obsessive-compulsive disorders, cortico-striato-thalamo-cortical feedback-loops are thought to be at the center of the dysfunctional network [55,59], while abnormal synchronization in highly similar neural populations is the source of Parkinsonian resting tremor [34,43].

At times, transitions between different pathological states or between states of health and disease are linear and balanced, at other times they are discontinuous and abrupt, such as in nonlinear phase transitions accompanied by critical fluctuations described by synergetics to occur in physical systems. Such transitions have been reported for unipolar or bipolar cyclic depression (e. g., [1]), and also for schizophrenia [68]. The usefulness of the concept of attractors in psychopathology is best reflected by the final common pathway of different disorders with similar phenomenology and syndromal patterns. Different initial conditions and different qualities and degrees of stressors and vulnerability factors may result in similar pathological end-states on the one hand. But on the other hand, small fluctuations within the intrapsychic or environmental conditions or small differences of some boundary or threshold conditions may result quite different disorders or may decide between health and disease (for a dynamical simulation of major depression see Schaub and Schiepek [56]). The encouraging message synergetics delivers is that while the structure of a generic system may stay unchanged, small changes in control parameters, threshold conditions, and internal or external fluctuations are able to trigger dramatic changes in the behavior of the system. As a consequence, therapy exerting changes of these parameters is thus able to trigger return of the system to a healthy state.

For illustrative purposes we present results of a computer simulation of different chronic courses of schizophrenia [57]. A qualitative network model of five macroscopic variables was transformed into a set of nonlinear difference equations, with each equation describing and determining the change rate of each variable from t to $t+1$. The empirical references of the simulation model were empirical studies of the chronic course of schizophrenic patterns, mostly mixed psychotic episodes, healthy functioning, and chronic states. For example, Ciompi and Müller [12] report on eight different patterns in the long-term evolution of schizophrenia, most of them reproduced by our model. These patterns result from various combinations of slow vs. acute onset, acute episodes vs. progressive deterioration, and remission vs. chronic end-state.

Variables taken into account were chosen from reviews in psychiatric and psychological schizophrenia research (e. g., [6,7,11]). Selected were (1) degree of cognitive disorders, (2) emotional and interpersonal stress, (3) withdrawal and social isolation, (4) degree of expressed (negative) emotions in the social environment of the patient, and (5) positive symptoms like delusions and hallucinations. The parameters mediating the nonlinear interplay of these macroscopic variables or order parameters were (a) diffuseness of affective-cognitive schemata as a central long-term vulnerability of mental functioning, (b) dopamine and serotonin metabolism, (c) social deficits and lack of competencies, (d) genetic risk for schizophrenia, and (e) some parameters mediating mixed feedback processes, especially the negative feedback responsible for antipsychotic damping effects of the pathology. Results of the simulation runs are indicated in Fig. 1. The simulation reproduces most precisely (a) episodic patterns with prodromal symptoms and acute onset, (b) acute onset, but continuous evolution with chronic end-state, and (c) slow and smooth onset with chronic long-term course (see also [76]).

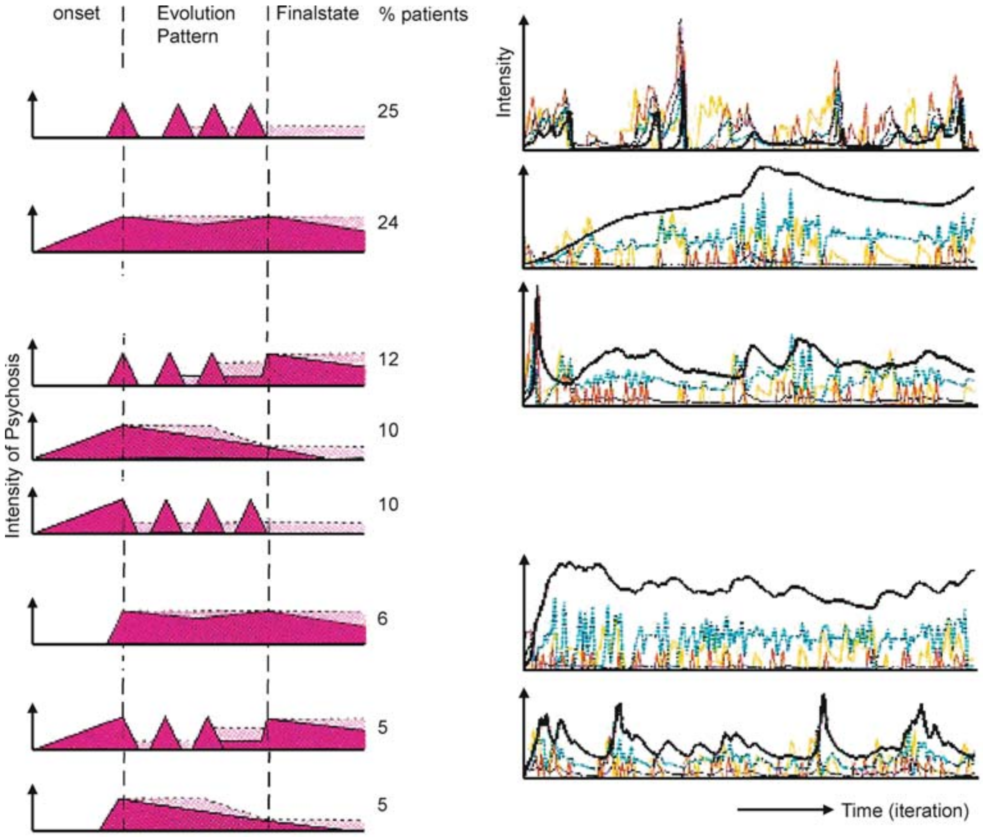
Other simulation models are effective on the microscopic level of neural networks. Kruse et al. [27] introduced a model focusing on the coupling dynamics between neurons processing brain correlates of social experiences. If unable to learn from cues delivered by the relevant environment, this system will fail to establish adaptive and coherent structures. When inducing fluctuations which promote re-learning and self-healing processes, the neural network causes incoherent and chaotic behavior. Most current models of schizophrenia take into account the neural circuits of relevant brain regions (cortical areas, basal ganglia like striatum and pallidum, thalamic areas, brain stem centers) and particularly the equilibria between different neurotransmitters and neuromodulators (e. g. [9]). The complicated local balances and their (non-) equilibria states are in the focus of the strongly evolving field of computational or systems neuroscience [15,44].

Not only central neuroscience has benefitted from concepts introduced by synergetics, however, but also physiology studying the effects of the autonomic nervous system (ANS) activities on peripheral systems, such as the cardiovascular, the respiratory, or the microcirculation system. Such activities are most prominent as the ANS engages in the mediation of emotions.

Self-Organized Synchronization Patterns in Peripheral Physiological Systems

For decades, the study of the ANS involvement in emotional arousal and its impact on the cardiovascular system has attracted clinical and scientific attention in psychology and psychophysiology (e. g. [65]). Ever since the seminal findings of W. B. Cannon [8] and H. Selye [63] on the general adaptation syndrome, colloquially condensed as stress, the clinical relevance of emotional responses became ultimately clear. This obvious clinical relevance is thwarted by the fact that direct observations of ANS activity in humans are restricted not only because of ethical constraints but also because of the fear to provoke what they strive to detect. Therefore, the study of the ANS in humans had to rely for a long time preferably on indirect measures, such as the power spectral density (PSD), a linear computation method. Based on the fast Fourier transform (FFT) which extracts periodic components in the frequency domain, the PSD was favored by many researchers due to its computational ease to analyze frequencies inherent in the two branches of the ANS, the parasympathetic (PNS) and the sympathetic nervous system (SNS). This allowed to divide the effects of the PNS and SNS activity on variations of the heart rate, the so-called heart rate variability (HRV), into three major variance components, the very low frequency (VLF) band below 0.04 Hz, the low frequency (LF) band between 0.04–0.15 Hz, and the high frequency (HF) band between 0.15–0.45 Hz. While the origin of both VLF and the HF bands is not debated, controversy reigns whether the origin of the LF band is attributable to SNS activity or whether it represents a mixture of SNS and PNS activity. Subsequently, some authors propose calculating the LF to HF ratio assumed to reflect the sympathovagal balance (for an overview see [73]). There has been growing discontent and criticism as to the validity of such drawer style classifications based on the consideration that the PSD, or FFT resp., as a linear routine is only able to detect linear properties, that are to some extent included in most physiological signals [23,86]. That, however, should restrict and limit its use since an increasing body of scientific evidence is demonstrating the obvious: In times of adaptation and rapid changes – a hallmark of life and its living systems – healthy ANS activity exhibits nonlinear dynamics necessary to mediate responses appropriate to those change processes. This is particularly the case for emoting as one of the most volatile change patterns.

However, this is not only true for discrete emotion transitions but also for a process crucial for the maintenance of health, namely psychophysical relaxation. Contrary to the rigid scheme depicted above, Perlitz and coworkers have introduced a relaxation model which takes into account adaptive, self-organizing characteristics of the central and peripheral subsystems involved in the psychophysical relaxation process. They scrutinized the physiological conditions and interactions observed with the emergence of a frequency at ca. 0.15 Hz, which in terms of the classical scheme is attributed to the transition between parasympathetic and sympathetic nervous activity. This frequency prevailed at different amplitudes in HRV, blood pressure and respiration, but foremost in the microcirculation of the forehead skin. Using several nonlinear methods, such as wavelet time frequency distribu-



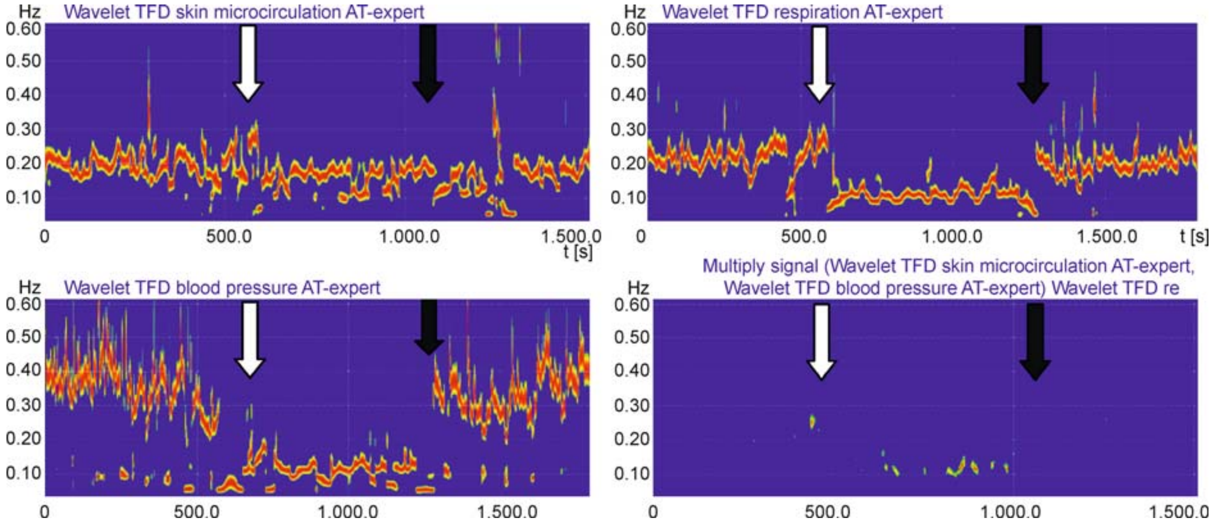
Self-Organization in Clinical Psychology, Figure 1
Different patterns of the long-term evolution of schizophrenia (empirical data from a study by Ciompi and Müller [12]) (*left*) were reproduced by simulations based on a set of five coupled nonlinear difference equations with different parameter values [57] (*right*)

tions (TFD) or post-event-scan (PES) analysis, this 0.15 Hz frequency band (range 0.12–0.18 Hz) emerged or erupted with amplified oscillations and periods of 6–7 s in all time series of subsystems under study. The emergence clearly depended on psychomotor drive reduction which can be either reduced by taking naive relaxation maneuvers (such as closing the eyes), or be enhanced using auto-suggestive means, such as autogenic training. Their zest to elaborate the origin of this frequency was supported by invasive observations with anesthetized dogs made by Lambertz and colleagues who had presented their findings earlier. They found a rhythm at a similar frequency which originated in reticular brainstem neurons of freely breathing dogs when administering narcotics to reduce drive. Followed by the emergence in those unspecific reticular neurons, this frequency also emerged in arterial blood pressure, HRV, and respiration [32,47,48]. This reticular rhythm, termed retR, was unaffected by changes in the frequency of respiration or arterial blood pressure which could both be presumed to exert distinct influences owing to linear models. Rather, in these experiments respiration and HRV were entrained to the 0.15 Hz band at 1:1, 2:1 and 1:2 integer number ratios which are, according to Bethe [5], an outflow of central-peripheral order–to–order transitions. With regard to parallels in frequency and dynamics observed in man and dog, Perlitz and coworkers suggested that also in humans the 0.15 Hz band most likely originates from reticular neurons of the lower brainstem network [46,47,48].

In summary, the findings presented in Fig. 2 underpin the theory of synergetics, since there is reason to regard the ca. 0.15 Hz frequency as an order parameter and the level of mental drive as control parameter. The ca. 0.15 Hz frequency is a prominent example of biological pattern formation lacking external or macroscopic control.

Nonlinear Dynamics in the Communication of Patient and Therapist

As mentioned above, psychotherapy is usually conceptualized as the application of psychological treatments to pa-



Self-Organization in Clinical Psychology, Figure 2

Wavelet time frequency distributions (TFD) of peripheral noninvasively obtained recordings of a female expert in autogenic training (AT, 56 yrs., healthy, non-smoker, 15 yrs practice AT). ***Top left*****: TFD of glabella skin microcirculation photoplethysmography;** ***top right*****: TFD of chest respiration related movements;** ***bottom left*****: TFD of peripheral systemic arterial blood pressure;** ***bottom right*****: So-called "joined TFD" of PPG-, respiration- and blood pressure-TFD, a novel method by Besting and colleagues (2005) (multiplying TFDs yielding only frequencies prominent at identical times and identical frequencies, www.Simplana.de) used to compute the intersection of TFD time series.** ***White arrows*** **mark the start of AT,** ***black arrows*** **mark the end of AT. In the TFD of PPG, the main frequency is at ca. 0.21 Hz prior to the onset of AT and is clearly stabilized at ca. 0.18 Hz with the start of AT, with signs of dissociation when terminating AT. The TFD of respiration supplies ample evidence of an order–order transition triggered by the practice of AT: The main frequency plummets from ca. 0.25 to 0.15 and 0.07 Hz to be maintained at ca. 0.12 Hz. With termination of AT, the main frequency skips back to frequencies shown beforehand. The TFD of systemic arterial blood pressure exhibits an intersection of approx. 90% during the AT section, but also few minor intersections before and after AT (data not shown); the joined TFD intersection shows merely few frequency "spots" at ca. 0.12 Hz during AT**

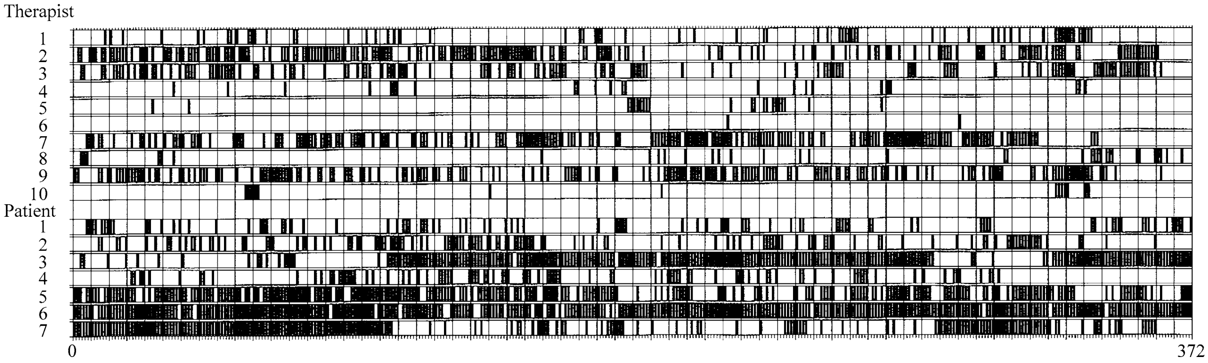
tients in order to change their problem states and diseases. However, as different research programs revealed during the last decades, psychological change processes show all important features of nonlinear systems – like deterministic chaos, nonstationary phase transitions, and nonlinear coupling between patient and therapist. Physiological synchronization appears to be realized at an interpersonal level (between therapist and patient) as well as between different phenomenological levels of the interpersonal system (speech qualities and psycho-physiological variables). In a study of Villmann et al. [83] heart rate, respiratory frequency, muscular tension, and skin conductivity were measured from both, therapist and patient, during 37 therapy sessions. Speech production was analyzed by the Mergenthaler model focusing on emotional feeling and cognitive referential activity/abstraction [38]. Physiological data were analyzed by an artificial neural network approach (growing self-organizing map), which uses a kernel smoothing for improved data density estimation. It was possible to generate an entropy model of psycho-physiological variability detecting emotionally instable phases during the therapy process. The entropy reflecting psycho-physiological and emotional variability was related to the dramatic value of speech analysis according to the cycle model of Mergenthaler.

Empirical evidence exists also for synchronized chaoto-chaotic phase transitions in the brains of therapist and patient during a therapeutic interview, measured by local largest Lyapunov exponents in the EEGs of both interaction partners [52].

Taking into account the importance of the therapeutic relationship for the treatment outcome the attention of a study realized by Schiepek and co-workers focused on the interactional process between therapist and patient [26,58]. The authors used the method of sequential plan analysis, which is a development of the hierarchical plan analysis proposed by Grawe and Caspar (e. g. [10]). Plans in this sense are verbally or non-verbally communicated intentions of self-presentation in a social situation. Patient's and therapist's interactional behavior was analyzed on the basis of video recordings. Two complete therapies (13 and 9 therapy sessions, resp.) were encoded with a sampling rate of 10 s (Fig. 3). The construction of an inclusive hierarchical plan-analysis leads to an

Self-Organization in Clinical Psychology, Table 1
Second-order plans and categories of self-presentations as identified by the hierarchical plan analysis of a complete 13-session psychotherapy. Encoding of therapist and patient. Plans and categories are used as ideographic observation categories for the sequential plan analysis

	Second-order plans	Categories of self-presentation
Therapist	1 show competence 2 encourage a trusting relationship 3 show understanding 4 motivate her	I encourage trust/create a secure atmosphere
	5 encourage her to reflect on her patterns of thinking 6 confront her with her avoidance and problem behavior	II confrontation/exposing to insecurity
	7 activate her 8 show her that she is responsible	III encourage self-responsibility of the patient
	9 guide her focus of attention 10 give her structure	IV activate structuring work
Patient	1 demonstrate strength and competence 2 make it clear that things are or have been difficult 3 be a good patient/create a good relationship to the therapist	I search for sympathy/appreciation/good relationship
	4 show that your suffering is strongly influenced by external causes 5 ask for help from the therapist	II externalization/demonstration of helplessness
	6 show interest and willingness in solving your problems 7 protect yourself from threatening changes	III problem-oriented work (self-relatedness vs. avoidance)

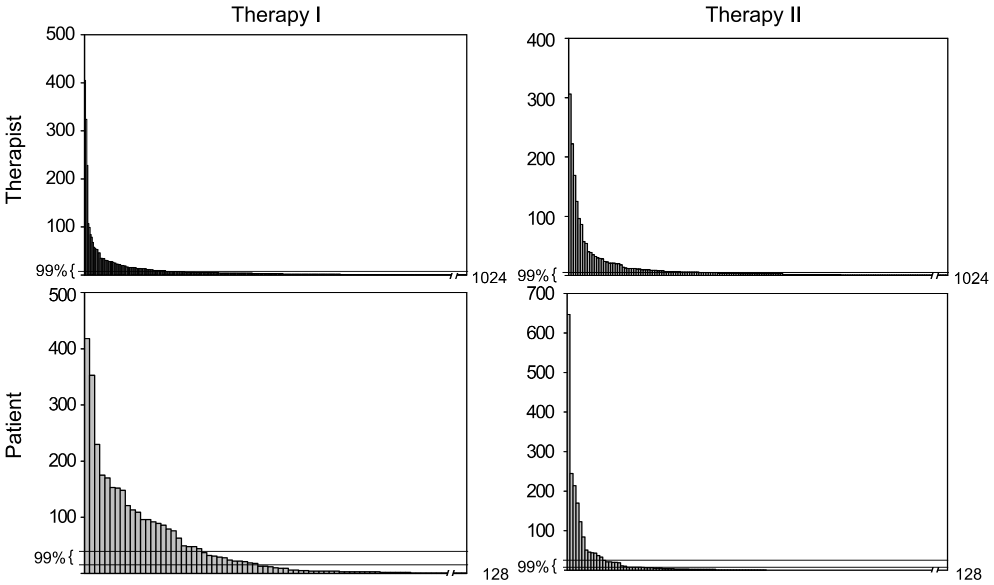


Self-Organization in Clinical Psychology, Figure 3
Nominal sequences of interactional plans of the therapist (*top*) and the patient (*bottom*) during a psychotherapy session. The sampling rate is 10 s. Different plans can be realized simultaneously. The pattern looks like a music score with the plans representing the different instruments of an orchestra. A sonification of the score of plans coded from a 13-session psychotherapy is recorded on a DVD added to the textbook of Haken and Schiepek [21]

ideographic categorical system for the observation of the client–therapist interaction (Table 1).

The first hints of order in the dynamics came from the distribution of simultaneous configurations (on-off-patterns) of plans in the scores. This distribution follows a power law ($1/f^a$) demonstrating a distinct structure order within the data (Fig. 4). Following Bak et al. [2], power law-distributions as demonstrated in Fig. 4 emerge from self-organized criticality within dynamic systems.

Further data analysis was based on the time series of the highest-level categories, the so-called categories of self-presentation (see Table 1). Since in the hierarchical system of the plan analysis the operators at the lowest observation level were quantified by intensity ratings, the plans and the self-presentation categories at the top level integrating the lower level categories were also quantified. The time series were analyzed by methods which are sensitive to the nonlinearity as well as the nonstationarity of the time series [21,26,58,70]. Nonlinearity was proofed by surrogate data tests [51] using random surrogates and FFT-based phase-randomized surrogates [69]. Whereas fractal dimensionalities of the empirical time series (based on



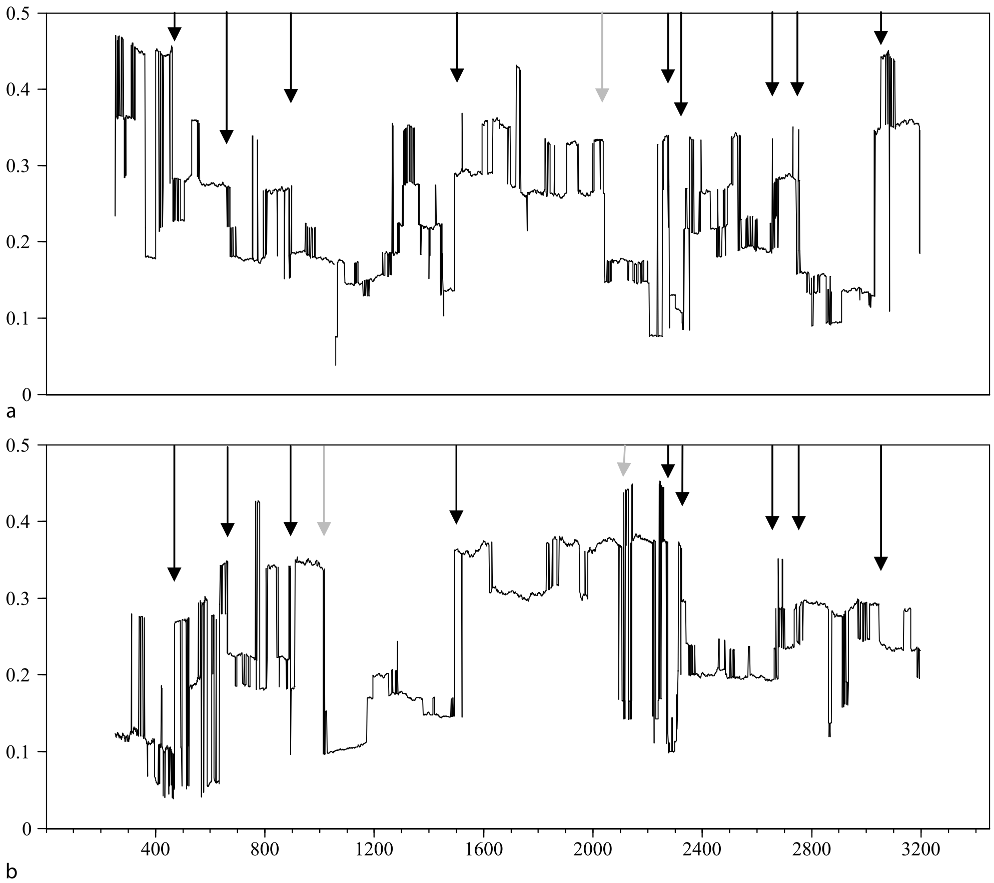
Self-Organization in Clinical Psychology, Figure 4

Empirical frequencies of constellations of interactional plans realized by therapists (10 plans) and patients (7 plans) within two psychotherapies (therapy I: 13 sessions, therapy II: 9 sessions). *X*-axis: Number of all possible configurations of plans (therapist: $2^{10} = 1024$, patient: $2^7 = 128$) ordered by the frequency of their realization. *Y*-axis: Frequencies of plan configurations. The distributions follow a power-law ($1/f^a$) distribution

the correlation dimension D2 as well as mean Pointwise D2 [66]) saturated at finite values (convergence to a fractal dimensionality of about 6), random and FFT-surrogates did not. The methods of PD2 [66] and of the local largest Lyapunov exponents (algorithm from Rosenstein et al. [53]) were used to identify phase-transition like discontinuities. Following the evolution of PD2 dimensionalities, both therapies realized nonstationarities, and both therapies showed periods of strongly synchronized (with correlations from 0.80 to 1.00) and anti-synchronized PD2-processes (with correlations from −0.80 to −1.00) between patient and therapist. Quite similar and even more pronounced dynamical jumps were to be seen in the development of the local largest Lyapunov exponents (Fig. 6), representing changes in the chaoticity of a time signal [26]. An important part of the discontinuities of the LLLE were exactly synchronized between patient and therapist. Obviously both persons create a dynamic self-organizing communication system, which allows for the individual change processes of the patient.

These results get support from nonlinear coupling measures between the time series of the interaction partners. Pointwise transinformation as well as pointwise coupling conditional divergence [33,80] were applied to the data, and both indicate changing and time-dependent coupling strengths between the time series of the interaction partners. There is no priority to the therapist's influence on the patient, which contradicts the classical idea that input from the therapist should determine the client's output. The other way round is also true and both constitute the circular causality of psychotherapeutic self-organization.

In other studies, sequential plan analysis was applied to the microdynamics of group interaction [21]. In a group of five persons a creativity and problem solving task was to be solved within 2,5 h (creation of ideas, rules, and physical handicraft realization of a prototype board game from different materials). Similar to the psychotherapy study the sampling rate was 10 s. The superordinate plans which could be identified for all five persons were (1) spontaneity and emotional engagement vs shyness, restricted behavior, and orientation to social norms, (2) engagement in the group interaction and in positive social climate, (3) task orientation. Length of time series was about 810 coding points (= intervals). D2 as well as mean PD2 estimates saturated at a fractal dimensionality of about 5 for all categories. The embedding of the time series was realized by two ways: (1) The phase space was constituted by the three dimensions of superordinated plans with five trajectories representing the five group members, or (2) the phase space was constituted by the five persons with three trajectories representing the time course of the three plans



Self-Organization in Clinical Psychology, Figure 5
Synchronized jumps in the dynamics of local largest Lyapunov exponents (*black arrows*). *Grey arrows* indicate not clearly synchronized changes. a Therapist, b Patient

Self-Organization in Clinical Psychology, Table 2
Factors (principal component analysis) of the Therapy Process Questionnaire (TPQ). Factor analysis was based on 94 therapy processes (mean stay = 66 days, daily ratings). Seven first-order factors (*right*) are related to three second-order factors (*left*). Numbers behind the first-order factors indicate factor loadings on second-order factors (for details see [21])

I(2) Change involvement	I Therapeutic progress/confidence in treatment effects/self-efficacy (.571) VI Intensity of therapeutic work/motivation to change (.596) V Opening of perspectives/personal innovations (.649)
II(2) Relationship/Social climate	III Quality of the therapeutic relationship/openness/confidence in the therapist (.705) II Ward atmosphere, social relationship to other inpatients (.692)
III(2) Emotionality	IV Dysphoric emotions/self-relatedness (.732) VII Impairment by symptoms and problems

(additional embedding dimensions result from time delay coordinates). In both cases PD2 results show an evolving pattern of quasi-attractors with changing complexity, and LLLEs (algorithm from Rosenstein et al. [53]) portray chaoto-chaotic phase transitions with clear-cut and interpersonally synchronized jumps – similar to the dyadic interaction of the psychotherapy study.

Self-Organization in Human Change Processes

A quite different approach to human change processes focuses on inpatient treatments at a hospital of psychosomatics. In a study by Schiepek and coworkers (results in [21]) 94 change processes were investigated, realized by 91 inpatients with different diagnoses (depression, anxiety disorders, posttraumatic stress disorders, eating disorders, somatoform disorders, and others). The time series data was produced by patients' self-ratings which were completed once a day in the evening. For this purpose a 53-item rating sheet was developed (*Therapy Process Questionnaire [TPQ]*, [21]) whose factor analysis resulted in a solution of seven factors defining the subscales of the questionnaire (Table 2). The ratings combined seven-step Likert scales and visual analogue scales especially for ratings of emotions. TPQ measurements reflect important aspects of the patient's experience of progress and goal attainment, emotional involvement, self-efficacy, therapeutic relationship, social relations with other inpatients, and the ward atmosphere.

The inclusive outcome criterion integrated the following measures: Inventory of Interpersonal Problems (IIP), Gießener Beschwerdebogen (GBB), Hospital Anxiety and Depression Scale (HADS), Questionnaire for Social Support (F-SOZU), a life-quality questionnaire (Münchener Lebensqualitäts-Dimensionenliste), a self-efficacy questionnaire (Fragebogen zur Generalisierten Kompetenzerwartung), the Sense-of-Coherence Questionnaire, and an interview-based assessment of personal resources. Additionally, therapists and patients scored the overall treatment effectiveness and treatment quality.

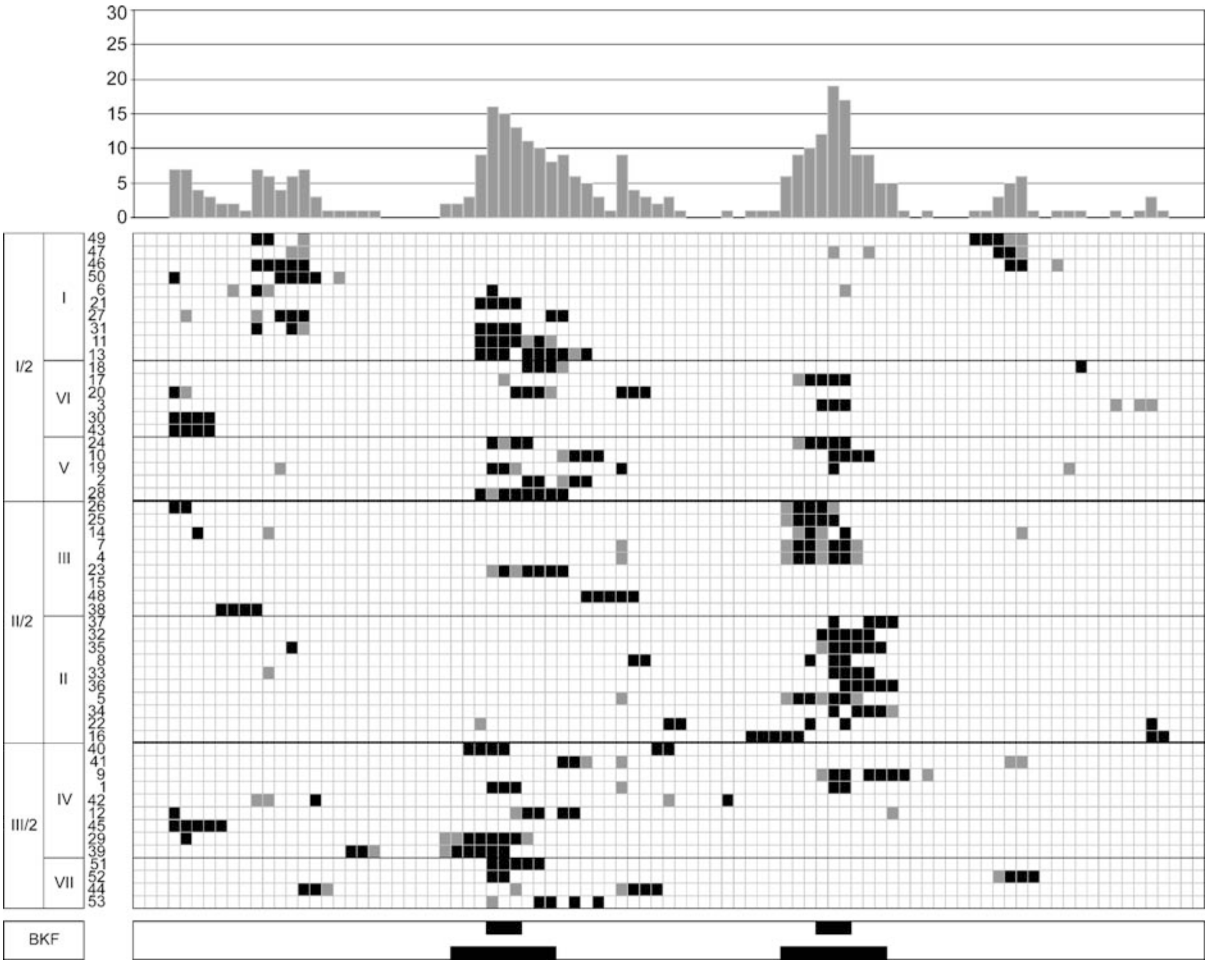
Results confirmed synergetic conceptualizations of how psychotherapy works and corroborated hypotheses drawn from this model. Here therapy is supposed to provide support for the patient's own self-organization processes, which should be characterized by cascades of order-to-order transitions accompanied by critical instabilities of the process. Pathological and restrictive order should be transformed into more flexible and adaptive patterns of behavior, and the synchronization of the different aspects of the patient's experience should undergo some transformations. Exactly this could be observed.

Significant correlations exist between the local maxima of critical fluctuations and the outcome of psychotherapy. The local maxima were defined by the difference between the mean dynamic complexity of the whole psychotherapy process and the maximum of the complexity which was observed during the process. Correlations were −0.455 (second-order factor I: "Change involvement" of the TPQ, $p = 0.002$), −0.431 (second-order factor 2: "Relationship/social climate", $p = 0.003$), and −0.572 (second-order factor 3: "Emotionality", $p = 0.000$) (compare Table 2). Negative correlations result from the fact that increased local maxima of dynamic complexity correspond to reduced problems, symptoms, and impairment.

The *dynamic complexity* combines a fluctuation index with a distribution index. The fluctuation index measures the frequency and amplitude of the change rates of a time series between the reversals of the development within a scanning window gliding over the whole time series. For analysis purpose a window width of seven measurement points (= days) was introduced. The distribution index measures the scattering of realized values within a given scanning window. The more scores are restricted to only narrow intervals of the available scale range, the smaller the distribution index becomes. The score of this index increases as the interval filled by the realized values grows. The algorithm solves the problem of value distribution independently of the scale resolution, the width of the scanning window, and of any combination of these parameters.

In order to answer the question if the observed intensities of dynamic complexity reach critical values, intra-item calibration procedures were used in order to define adequate thresholds fitting to the actual dynamics. The time series of dynamic complexity were standardized by z-transformations, providing significance thresholds of 5% or 1%. Applying this threshold method to all items of the TPQ reduces the quantitative complexity signals of each time series to a three-step signal (not significant, complexity exceeds a 5% threshold, complexity exceeds a 1% threshold). A synopsis of these qualitative signals referring to all items of the TPQ gives an impression of the localization of critical fluctuations during the whole process. Dynamic complexities seem to be synchronized over many items and factors of the TPQ, resulting in the structure of columns of grey (<5%) or black (<1%) dots. In a large part of the investigated therapies such column-like structures could be identified. In an item-by-time synopsis they indicate phases of intensified as well as synchronized fluctuations and entropies of quite different aspects of the process. Consequently, these item-by-time synopses are called *complexity resonance diagrams* (Fig. 6).

In order to confirm the structures found within the complexity-resonance-diagrams, surrogate tests were realized based on random as well as on FFT-based surrogates of the time series. The empirical patterns are impressively different from the surrogate-based patterns (all realized comparisons with $p = 0.000$). Further support for phase-transition like phenomena in the change processes came from recurrence plots representing similarities and dissimilarities of dynamic segments of a whole time series [13,80,85]. This method is based on the embedding of time series into a phase space constructed by time-delay coordinates, a method which is also crucial in the algorithms for the estimation of dimensional complexity or chaoticity (e. g., Kolmogorov–Sinai-Entropy, Lyapunov Exponents). Neighbors in the time-delay phase space rep-

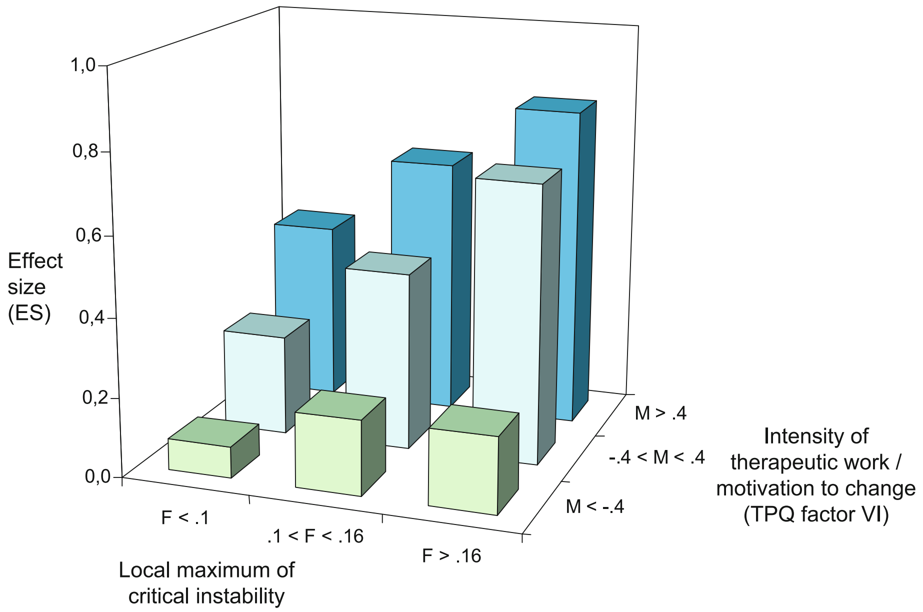


Self-Organization in Clinical Psychology, Figure 6

Complexity resonance diagram of a psychotherapy process. Such diagrams portray the threshold exceeding dynamic complexities of a process encoded by the 53 items of the Therapy Process Questionnaire (TPQ). *Gray dots*: 5% threshold of significance; *black dots*: 1% threshold of significance. *X*-axis: Days of hospital stay, *Y*-axis: Items of the TPQ arranged by the order of the factors as reported in Table 2. Window width for the calculation of dynamic complexities is 7. Column-like structures indicate phases of critical instabilities during the process

resent similar dynamic segments and are plotted by a dot in the recurrence plot. Dissimilarities are represented by empty columns in the recurrence plots, which in many cases exactly correspond to the columns of dots in the complexity-resonance diagrams. The overall correlation is -0.45, if small shifts (lags of $+$ or -3 measurement points at maximum) will be allowed. This means that periods of critical instability correspond to transient dynamics outside of the quasi-attractors established by the self-organizing system under consideration. These different ways to identify critical phase transitions are further validated by the time frequency distribution (TFD) of the time series. The TFD method uses wavelet spectra in order to scan the evolution of the frequency distributions within a signal [33,80]. It is a dynamic counterpart to the static fast Fourier transformation and allows for the identification of pronounced frequency amplitudes or changes in the frequency distributions. In the data set of the referred study these often appear exactly during the phase transitions which can be identified by other methods (see the synoptical representations of different time series analysis methods on the DVD in the textbook of Haken and Schiepek [21]).

An overall result of the study is shown in Fig. 7. It portrays the evidence that in order to bring forth change processes within self-organizing systems at least two conditions should be realized. The first condition: The degree of the control parameter energizing the system and push-

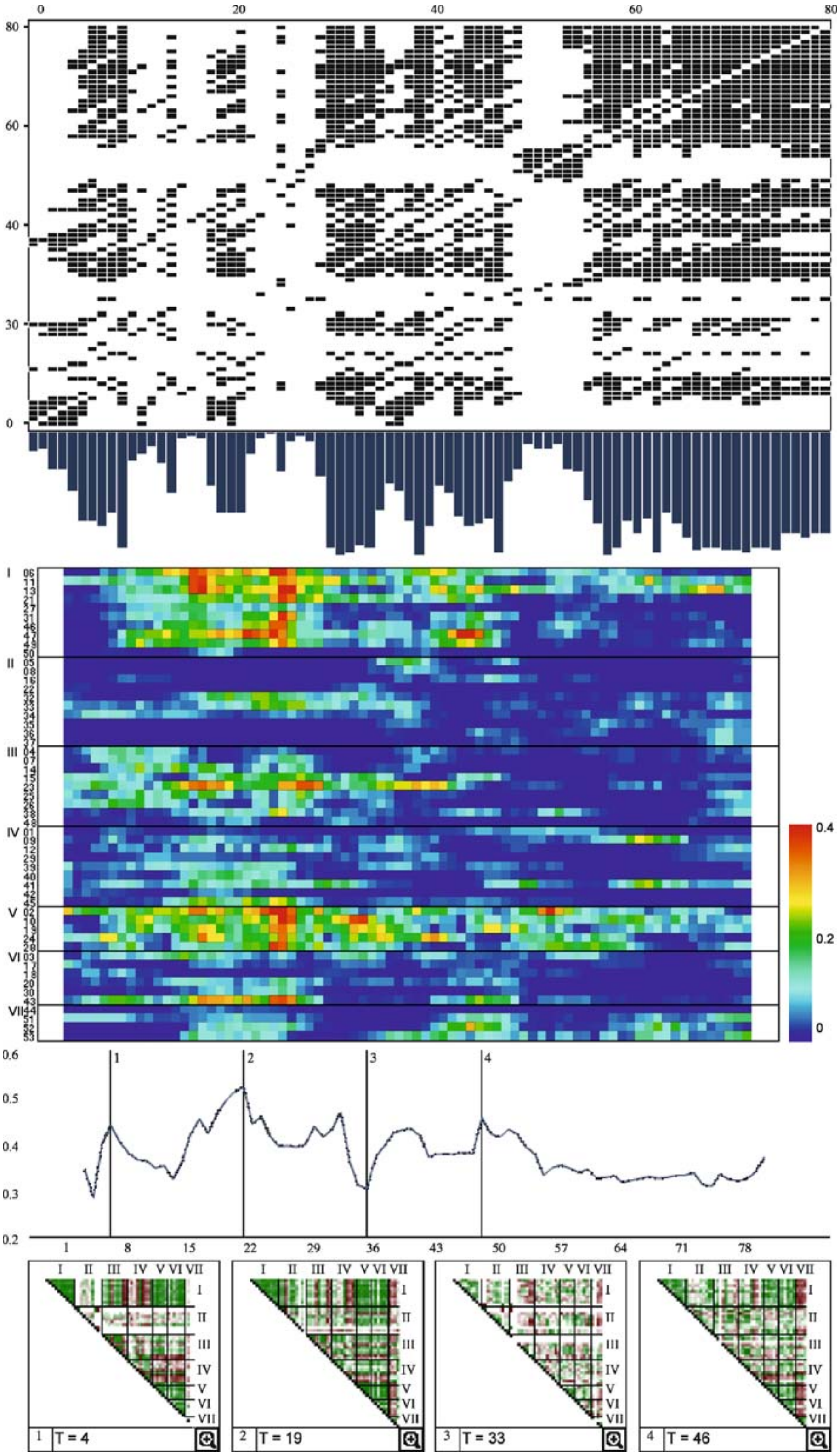


Self-Organization in Clinical Psychology, Figure 7
The effect size (ES) (mean ES of all outcome measures introduced in the study, see text) of inpatient psychotherapy is produced by an interaction between the local maximum of critical fluctuations and the intensity of the control parameter realized during the change process. The local maxima of fluctuations were defined by the difference between the mean dynamic complexity of the whole therapy process and the maximum of the complexity observed during the process. The diagram is based on the mean of the local maxima of all items. The control parameter was defined by the overall mean of the TPQ factor VI: Intensity of therapeutic work/motivation to change

ing it away from its actual equilibrium state should exceed a certain intensity level. With respect to psychotherapies this control parameter could be the patient's motivation to change including his engagement into the therapeutic work. Second condition: The degree of instability the system attains during its change process. This instability during emerging symmetries and symmetry breaking transitions is given by the local maximum of dynamic complexity during the hospital stay. The interaction of both conditions results in treatment effectiveness. A third important condition is not represented in Fig. 7: It is the experienced stability of the outer environment (context at the ward or therapeutic bond) or of the inner environment (as self-esteem, self-confidence, or activated resources). This context of stability is a prerequisite for a system to undergo critical instabilities.

The Concept of Self-Organization Promotes New Information Technologies in Clinical Psychology – The Synergetic Navigation System

Since self-organization and nonlinear dynamics seem to be ubiquitous in human change processes, it should be helpful to go beyond the diagnostics of steady states to an assessment of dynamics. Practitioners should get information on the therapy and its features *during* the ongoing process in order to use this information for an adequate placement of interventions and a control of the dynamics. "Controlling" self-organization processes in psychotherapies means the generation and co-creation (together with the patient) of adequate boundary conditions, the decision to do or to retain certain interventions, and to support the dynamics which the system is creating by itself. The patient takes an active and cooperative role in this understanding of data-based and co-creative change processes. Another important motivation for the development of real-time assessment comes from the evidence that most of the empirically identified specific and non-specific factors driving therapeutic change processes are connected with specific persons (the concrete therapist who meets a concrete patient in a concrete setting) and evolve by its nonlinear interactions in specific systems. These factors are (i) personal features of the patient like his motivation to change, his premorbid adaptation and degree of social functioning, personality integration, ego-strength, or comorbidities, (ii) personal and professional features of the therapist like his own personality integration, social and professional competencies, allegiance to his approach of


0
20
40
60
80
80
60
40
30
0
I 06 11 13 21 27 31 46 47 49 50
II 05 08 16 22 32 33 34 35 36 37
III 04 07 14 15 23 25 26 38 48
IV 01 09 12 29 39 40 41 42 45
V 02 10 19 24 28
VI 03 17 18 20 30 43
VII 44 51 52 53
0.4
0.2
0
0.6
0.5
0.4
0.3
0.2
1
2
3
4
1
8
15
22
29
36
43
50
57
64
71
78
I II III IV V VI VII
1 T = 4
2 T = 19
3 T = 33
4 T = 46

◀ Self-Organization in Clinical Psychology, Figure 8

Synopsis of a psychotherapy process as monitored by the Synergetic Navigation System. The time course of the inpatient treatment of a patient with eating disorders portrays a clear cut phase-transition associated with critical instabilities. *Top*: Recurrence plot of the item "Today I was successful to do steps towards my personal goals". *Dots* represent recurrent segments of the time series, empty spaces represent transitions. *Middle*: Complexity resonance diagram of all items of the TPQ. Different from Fig. 6, the intensities of the dynamic complexity of each item is transformed into colors. Items are arranged by the order of the first- and second-order factors of TPQ. *Bottom*: Mean of all inter-item correlations irrespective of the sign (absolute values). This is a measure of the overall synchronisation of the patient's experiences as represented by the items of the TPQ. The correlation structure is shown at four measurement points (days) of the psychotherapy process ($t = 4$, $t = 19$, $t = 33$, $t = 46$). Intensity of green represents positive correlations, intensities of red represents negative correlations

doing therapy, stress-resistance, and so on, and (iii) factors of the professional and social context (see the so-called generic model of psychotherapy [29,42]). In consequence, *evidence-based treatments* should be based on the evidence of concrete data mirroring the ongoing change process and on the professional decisions reflecting this insight.

Real-time monitoring actually uses internet-based presentations (including PDA or cell phone technology) of outcome and process questionnaires. Data are sent to a server, where they are stored and analyzed. Professionals and patients can inspect the results whenever they want. Experiences with real-time feedback to therapists (based on an outcome questionnaire the patient fills out during the therapy sessions in an ambulatory or outpatient context) are encouraging. Lambert and co-workers (e. g., [31]) were able to identify processes on the way of getting difficult or unsuccessful ("not on track" therapies, compared to more promising "on track" therapies), and helped therapists to correct these not-on-track dynamics by specific interventions. By this, threatening drop-outs could be avoided, bad results could be corrected, and on-track processes could be optimized and even shortened.

More sophisticated than the distinction between "on-track" and "not-on-track" courses is the feedback on self-organization features realized by a system based on synergetics [21]. The *Synergetic Navigation System* uses the therapy process questionnaire for daily ratings and applies methods from nonlinear time series analysis in order to identify important qualities of the change process. This are:

- Stability or instability of the dynamics as represented by the subscales (factors) of the TPQ (see Table 2), which is measured by the dynamic complexity
- Recurrence plots indicating transitions or repeating patterns
- Intensity of synchronization and time-dependent synchronization patterns between the items and the factors of the TPQ (realized by the cross-correlations of all items of the TPQ, calculated within a running window).

Figure 8 shows a synopsis of these analysis methods applied to a specific change process. Preceding the inspection of all analysis results the raw data series of the items and the time courses of the factors (z-transformed values) are available. Additionally patients can write an electronic diary after filling out the questionnaire. The diary entries can be presented within a gliding tip-tool running over the time series. By this, corresponding qualitative and quantitative information completes the picture.

The Self-Organizing Brain

The human brain is one of the most outstanding examples of a complex nonlinear system producing self-organized patterns of functioning. Since function corresponds to structure and vice versa, structural changes (changes of intersynaptic coupling strengths and network configurations, (re-)wiring patterns following the synchronized co-activity of neurons) can be explained by functional self-organization of neural populations. Perception, action and transition of action patterns, decision making, and cognitive, behavioral, as well as emotional learning are psychological functions following the principles of self-organization [21]. At a neural level they correspond to and are based on nonlinear brain dynamics. The emergence of order parameters and the occurrence of phase transitions can be described and measured on a psychological as well as on a neural level.

One of the phenomena modeled by synergetics is Gestalt perception – the construction of percepts and the switching of ambiguous visual patterns (e. g., Necker cube or stroboscopic alternative motion). These processes of Gestalt perception constitute the link between Gestalt psychology and actual mathematical modeling in synergetics [17]. The binding of different perceptual features or components to coherent structures or "qualia" seems to be due to synchronization processes of extended brain regions and converging integrative areas [64]. Pattern perception corresponds to pattern formation – as H. Haken puts it into pointed words. Tallon-Baudry et al. [71,72] measured enhanced gamma-band activity (30–50 Hz) in the EEG of the primary and secondary visual cortex while subjects identified a triangle within the offered stimulus material. This could be a fingerprint of corresponding neural synchronization processes. This activity occurred when subjects saw a real object (triangle) as well as a figural illusion of the object (Kanizsa triangle), but not if geometrical components could not be composed to a true Gestalt. The research group of Basar–Eroglu and Stadler [4] measured significant gamma-band activity in EEG during states of perceptual switching triggered by stroboscopic alternative motions. In summary: Perception of multistability is one of the multifold cognitive processes giving rise to 40 Hz enhancement in the cortex, and coherent oscillations reflect an important mechanism of feature linking in the visual cortex which corresponds to the emergence of a neural order parameter. Changing order parameter dynamics during different cognitive activities was shown by Schupp et al. [62]. Mental imagery of an object could be differentiated from its concrete perception. The dimensional complexity of prefrontal EEG was increased during sensory imagery compared to the real perception of the same object (compare [36]).

The well-known movement coordination paradigm modeled by Haken et al. [22] was used to demonstrate neural correlates of instability and symmetry breaking processes in the motor brain. The order parameter in this finger movement experiment is the relative phase of the index fingers of both hands. Metronome-pacing – with movement frequency as the control parameter – triggers the system from parallel (out-of-phase) to mirror (in-phase) movement. Meyer-Lindenberg et al. [39] showed that the emergence of patterns in open, nonequilibrium systems like the brain is governed by their stability in response to small disturbances. Transitions could be elicited by interference at the neural level. Functional neuroimaging (PET) identified premotor (PMA) and supplementary motor (SMA) cortices as having neural activity linked to the degree of behavioral instability, induced by increasing frequency of the finger movement. These regions then were transiently disturbed with graded transcranial magnetic stimulation (TMS), which caused sustained and macroscopic behavioral transitions from the less stable out-of-phase to the stable in-phase movement, whereas the stable pattern could not be affected. Moreover, the strength of the disturbance needed (a measure of neural stability) was linked to the degree of the control parameter (movement frequency) and thereby to the behavioral stability of the system.

Synergetic research in clinical psychology is now reaching the brain level. The aim of an actual fMRI-study [60] is the investigation of phase transitions of brain activity and related subjective experiences of patients during their psychotherapy process. Repeated fMRI scans are related to the degree of stability or instability of the ongoing dynamics (measured by the dynamic complexity of daily TPQ-ratings) as well as to the therapy outcome. Real-time monitoring by the *Synergetic Navigation System* allows for the identification of stable or unstable periods and by this for a decision on the appropriate moments of fMRI acquisitions. Three or four scans are realized during each of the psychotherapy processes of 15 patients. The study includes only patients with obsessive-compulsive disorder (OCD) of the washing/contamination fear subtype (DSM IV: 300.3), without any medication or comorbid psychiatric or somatic diagnoses. Patients are matched to healthy controls. (This research is a multi-center study of the Ludwig-Maximilians-University Munich, Institute of Psychology (Prof. Dr. Günter Schiepek, head of the project), and Clinic of Psychiatry (PD Dr. Oliver Pogarell, Dipl. Psych. Susanne Karch, Dr. Christoph Mulert), Hospital of Psychosomatic Medicine Windach/Ammersee and Day Treatment Centre Munich/Westend (Dr. Igor Tominschek, cand. Psych. Stephan Heinzel, Prof. Dr. Michael Zaudig), University Hospital Vienna/Astria, Clinic of Psychiatry (Prof. Dr. Martin Aigner, Prof. Dr. Gerhard Lenz, cand. med. Markus Dold, Dr. Annemarie Unger), MR Centre of Excellence, Medical University Vienna/Austria (Prof. Dr. Ewald Moser, Dr. Christian Windischberger).

OCD seems to be an appropriate model system for synergetic studies in clinical psychology, since the pathological order parameter is phenomenologically quite evident, the disease has an obvious and quite stable time course, and therapeutic phase transitions – if they do occur at all – are easy to be observed. OCD-specific functional neuroanatomy is partially known: Friedlander and Desrocher [14] report on an executive dysfunction model corresponding to the cortico-striato-thalamo-cortical feedback-loops involved in perseverations and compulsions, and on a modulatory control model involved in the pathological mechanisms of anxiety and distress provoking obsessions.

The visual stimulation paradigm of the study uses symptom provoking, disgust provoking, and neutral pictures. The disgust and the neutral pictures are taken from the International Affective Picture System, whereas the OCD-related pictures are photographed in the home setting of the patients, showing specific and individualized symptom provoking stimuli.

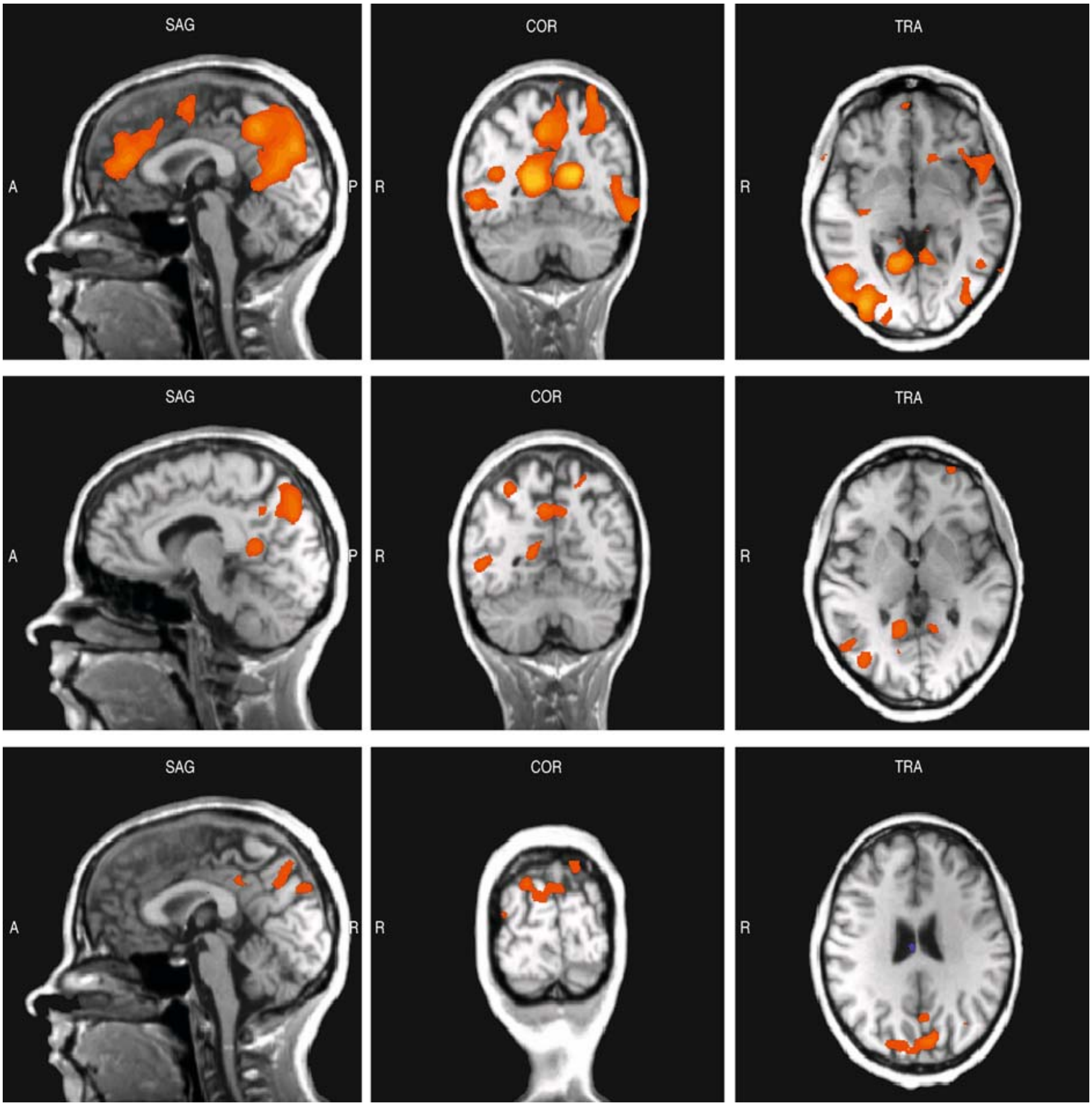
For illustrative purposes we report on the results of a single case. It is a female patient, whose fMRI scans were taken three times during the 59 days of their hospital stay at days 9, 30, and 57. The healthy control was also scanned three times at identical time intervals as the patient. The second acquisition was done after an intensive period of critical instability of the TPQ-based time series, but just before the flooding was started. (Flooding or response prevention is an essential therapy technique in the treatment of OCD, where patients are confronted with symptom provoking stimuli but abstain from performing compulsive rituals.) The instability of the patient's process was the precursor of an important personal decision to divorce from her husband. (It should be noted that the development of her OCD symptoms was in the context of a long-lasting marital conflict.) This decision was the essential phase transition of the therapy.

Indeed, the most pronounced changes in brain activity occurred from the first to the second fMRI scan, whereas BOLD response differences from the second to the third session were only slight. They perhaps represent the neural correlates of an important personal phase transition related to the resolution of a severe personal conflict. Because these changes occurred before the flooding procedure was started, this can be seen as indicator of an early rapid response in the therapy [29]. Additionally, marked alternations in brain activity were to be observed before or during symptom reduction took place (measured by the Y-BOCS), not afterwards.

Alternations in brain activity involved widespread areas, e. g. the medial frontal brain regions including anterior cingulate cortex, superior and middle frontal gyrus, inferior frontal and precentral gyrus, superior temporal gyrus, superior parietal lobe, cuneus, thalamus and caudate nucleus in both hemispheres, as well as the right fusiform gyrus (see Fig. 9 for a OCD to disgust contrast). Thalamic and basal ganglia activation is part of the dorsolateral-caudate-striatum-thalamus circuitry of OCD. Especially the caudate nucleus takes a role within the executive dysfunction model of compulsions, and its activity has been found to be reduced after treatment [40].

The function of the anterior cingulate cortex is interesting with regard to synergetics. The cingulate cortex comprises various functions like somatosensoric integration, mediation of affective and cognitive processes, control of attention, and processing of painful stimuli. Additionally, it plays an important role as conflict monitoring system: It is sensitive to ambiguous or conflicting information [81,82], is involved in decision making [24,54], and its activation is predictive to treatment outcome in depression (e. g. [37]). This is true especially for the dorsal (cognitive) structures of the ACC. It could be an indicator of symmetry states of brain functioning, which is characterized by two or more dynamic patterns or attractors in competition. In the present case, the ACC activation at the beginning of the therapy could be either part of the pathology or could be indicative for the critical instability of the cognitive-affective system of the patient, preparing her important decision. The second fMRI measure was conducted during a local minimum of critical fluctuations. If the impressive change in cingulate activation could be attributed to a changed critical symmetry state of the neural self-organization before vs. after the phase-transition or to changes in symptom severity cannot be decided within a single case study, but seems to be an interesting question to further research. Perhaps the fact that during the second fMRI measure the Y-BOCS score was nearly on the same level as during the first measure – only 14% reduction, compared to 50% reduction in dynamic complexity – could be a first argument in favor of the instability hypothesis.

The paradigm of self-organization is a very promising approach to clinical as well as other fields of psychology. Its interdisciplinary is due to the fact that the laws and principles of self-organization are true for neural, mental, and behavioral processes (and the corresponding data qualities). Interdisciplinary cooperation is underpinned by the unifying terminology as well as by the unifying formalism and modeling tools of synergetics. This opens new perspectives for basic and applied research, but also for the treatment of mental disorders. New developments in the real-time monitoring of human change processes based on synergetics and nonlinear science have been mentioned. Another field of encouraging developments is deep brain stimulation (DBS), which apply to neurological diseases as Parkinsonian or essential tremor, but also to psychiatric disorders as OCD or mayor depression [75]. The difference between new technologies (applying the mathematical instruments and concepts of synergetics as well as methods from stochastic phase resetting) and classical electrical deep brain stimulation is that normal DBS at high frequencies has a blocking effect on the stimulated target and mimics the effect of tissue lesioning. New technologies are demand-controlled, working with low stimulation frequencies, and avoid the suppression of neurons' firing. Its effect is a desynchronization of pathologically synchronized populations of neurons, using multi-site coordinated reset stimulation [74] or nonlinear delayed feedback stimulation [50]. Both methods counteract abnormal interactions and detune the macroscopic frequency of the collective oscillators – that is the abnormally established order parameters of neural synchronization. Thereby they restore the natural frequencies of the individual oscillatory units. Neurons get in the range of physiological functioning and can engage in changing and varying synchronization patterns. If altered synchronization patterns also change the coupling strength connecting synapses, a rewiring of neural nets could be reached. Changed function triggers the emergence of healthy at-



Self-Organization in Clinical Psychology, Figure 9

Brain activation patterns of a patient with OCD during psychotherapy. BOLD signals from a 1.5 Tesla fMRI scanner. *Top*: First scan (9th day of hospital stay; $x = 0$, $y = -55$, $z = -2$; p(uncor) <0.001). *Middle*: Second scan (30th day of hospital stay; $x = 8$, $y = -54$, $z = 5$; p(uncor) $<$ 0.001). *Bottom*: The third scan (57th day of hospital stay; $x = 0$, $y = -85$, $z = 26$; p(uncor) $<$ 0.001). Activations during the presentation of OCD-related pictures compared to activations during the presentation of neutral pictures (OCD > disgust)

tractors and by this changes the structure of neural networks. Perhaps in the future technologies of DBS or even non-invasive brain stimulation could be combined with psychotherapy and psychological navigation instruments developed to optimize self-organizing change processes.

Future Directions

The future developments of self-organization and complexity research in clinical psychology and psychotherapy will be interconnected to its acceptance in practice and training. Perhaps this sounds paradoxically, since in most other scientific fields the future depends on the investigations to basic research and to new technologies. Of course this holds also for synergetics and its applications to clinical psychology. However, it should be noted that complexity research and nonlinear dynamics are done since more than two decades in European academic psychology with poor impact to mainstream science. So, the future will depend on a greater number of new arriving and highly qualified students in this topic who do not avoid the touch with mathematics. Self-organization and complexity research including its mathematical backgrounds should become part of the training curricula in psychology and psychotherapy. Since the Synergetic Navigation System waits for its broad application in clinical and psychotherapeutic practice, a new decade of practice-based research can be started. But these developments depend on its acceptance by practitioners because of the competencies required for the widespread use of sophisticated methods. This integration of science with practice will open huge sources and new dimensions of data gathering on dynamic systems. An important database for outcome and time series data (including biomarkers) of human change processes is actually prepared.

Another stream of development is concerning the integration of psychological and biological/physiological data. Since human self-organization takes place on synchronized mental, social, and biological system levels, all of them should be taken into account in further research. One research paradigm was suggested in this chapter: The investigation of individual and social processes by the Synergetic Navigation System, and in parallel repeated brain scans using fMRI technology or other methods to get insight into brain dynamics (EEG, gene expression markers [25], immune or endocrine markers [61], or others). Two final remarks: First, future developments of synergetic-based minimal invasive DBS could be combined with psychotherapy and psychological interventions – as pharmacological and psychological treatments are combined nowadays. Second, the nonlinear networks underlying psychological as well as neural self-organization will not be understood without applying appropriate mathematical tools, giving raise to a new systemic psychology and neuroscience.

Bibliography

1. an der Heiden U (1992) Chaos in health and disease – phenomenology and theory. In: Tschacher W, Schiepek G, Brunner EJ (eds) Self-organization and clinical psychology. Springer, Berlin, pp 55–87
2. Bak P, Chen K, Creutz M (1989) Self-organized criticality and the 'Game of Life'. Nature 342:780–782
3. Basar E, Flohr H, Haken H, Mandell AJ (1983) Synergetics of the brain. Springer Series in Synergetics, vol 23. Springer, Berlin
4. Basar-Eroglu C, Strüber D, Kruse P, Basar E, Stadler M (1996) Frontal gamma-band enhancement during multistable visual perception. Int J Psychophysiol 24:113–125
5. Bethe A (1940) Die biologischen Rhythmusvorgänge als selbständige und erzwungene Kippvorgänge betrachtet. Pflügers Arch 244:1–42
6. Böker W, Brenner HD (1996) Stand systemischer Modellvorstellungen zur Schizophrenie und Implikationen für die Therapieforschung. In: Böker W, Brenner HD, Genner RM (eds) Integrative Therapie der Schizophrenie. Huber, Bern, pp 17–32
7. Böker W, Brenner HD (eds) (1989) Schizophrenie als systemische Störung. Huber, Bern
8. Cannon WB (1915) Bodily changes in pain, hunger, fear, and rage: An account of recent researches into the function of emotional excitement. Appleton, New York
9. Carlsson A (2006) The neurochemical circuitry of schizophrenia. Pharmacopsychiatry 39(Suppl 1):S10–S14
10. Caspar F (1996) Beziehungen und Probleme verstehen. Eine Einführung in die psychotherapeutische Plananalyse. Huber, Bern
11. Ciompi L (1989) Zur Dynamik komplexer biologisch-psychosozialer Systeme: Vier fundamentale Mediatoren in der Langzeitentwicklung der Schizophrenie. In: Böker W, Brenner HD (eds) Schizophrenie als systemische Störung. Huber, Bern, pp 27–38
12. Ciompi L, Müller C (1976) Lebensweg und Alter der Schizophrenen. Springer, Berlin
13. Eckmann JP, Oliffson Kamphorst S, Ruelle D (1987) Recurrence plots of dynamical systems. Europhys Lett 4:973–977
14. Friedlander L, Desrocher M (2006) Neuroimaging studies of obsessive-compulsive disorder in adults and children. Clin Psychol Rev 26:32–49
15. Friston KJ, Harrison L, Penny WD (2003) Dynamic causal modelling. NeuroImage 19:1273–1302
16. Haken H (1990) Synergetics – an introduction. Nonequilibrium phase transitions in physics, chemistry, and biology. Springer, Berlin (first edition 1977)
17. Haken H (1990) Synergetics as a tool for the conceptualization and mathematization of cognition and behavior – How far can we go? In: Haken H, Stadler M (eds) Synergetics of cognition. Springer, Berlin, pp 2–31
18. Haken H (1996) Principles of brain functioning. A synergetic approach to brain activity, behavior, and cognition. Springer, Berlin
19. Haken H (2002) Brain dynamics. Springer, Berlin
20. Haken H (2004) Synergetics. Introduction and advanced topics. Springer, Berlin
21. Haken H, Schiepek G (2006) Synergetik in der Psychologie. Selbstorganisation verstehen und gestalten. Hogrefe, Göttingen
22. Haken H, Kelso JAS, Bunz H (1985) A theoretical model of phase transition in human hand movements. Biol Cybern 51:347–356
23. Kettunen J, Keltikangas-Järvinen L (2001) Intraindividual analysis of instantaneous heart rate variability. Psychophysiology 38:659–668

24. King-Casas B, Tomlin D, Anen C, Camerer CF, Quartz SR, Montague PR (2005) Getting to know you: Reputation and trust in a two-person economic exchange. Science 308:78–83
25. Koch JM, Kell S, Hinze-Selch D, Aldenhoff JB (2002) Changes in CREB-phosphorylation during recovery from major depression. J Psychiat Res 36:369–375
26. Kowalik ZJ, Schiepek G, Kumpf K, Roberts LE, Elbert T (1997) Psychotherapy as a chaotic process II: The application of nonlinear analysis methods on quasi time series of the client-therapist-interaction: A nonstationary approach. Psychother Res 7:197–218
27. Kruse P, Carmesin HO, Stadler M (1997) Schizophrenie als Korrespondenzproblem plastischer neuronaler Netze. In: Schiepek G, Tschacher W (eds) Selbstorganisation in Psychologie und Psychiatrie. Vieweg, Braunschweig, pp 171–190
28. Köhler W (1947) Gestalt psychology. Liveright, New York
29. Lambert MJ, Ogles BM (2004) The efficacy and effectiveness of psychotherapy. In: Lambert MJ (ed) Bergin and Garfield's handbook of psychotherapy and behavior change. Wiley, New York, pp 139–193
30. Lambert MJ, Whipple JL, Smart DW, Vermeersch DA, Nielsen SL, Hawkins EJ (2001) The effects of providing therapists with feedback on patient progress during psychotherapy: Are outcomes enhanced? Psychother Res 11:49–68
31. Lambert MJ, Whipple JL, Vermeersch DA, Smart DW, Hawkins EJ, Nielsen SL, Goates M (2002) Enhancing psychotherapy outcomes via providing feedback on client progress: A replication. Clin Psychol Psychother 9:91–103
32. Lambertz M, Vandenhouten R, Grebe R, Langhorst P (2000) Phase transitions in the common brainstem and related systems investigated by nonstationary time series analysis. J Autonom Nerv Syst 78:141–157
33. Lambertz M, Vandenhouten R, Langhorst P (2003) Transiente Kopplungen von Hirnstammneuronen mit Atmung, Herzkreislaufsystem und EEG: Ihre Bedeutung für Ordnungsübergänge in der Psychotherapie. In: Schiepek G (ed) Neurobiologie der Psychotherapie. Schattauer, Stuttgart, pp 302–324
34. Lenz F, Kwan H, Martin R, Tasker R, Dostrovsky J, Lenz Y (1994) Single unit analysis of the human ventral thalamic nuclear group. Tremor-related activity in functionally identified cells. Brain 117:531–543
35. Lewin K (1951) Field theory in social psychology. Harper, New York
36. Lutzenberger W, Elbert T, Birbaumer N, Ray WJ, Schupp H (1992) The scalp distribution of the fractal dimension of the EEG and its variation with mental tasks. Brain Topogr 5:27–34
37. Mayberg HS, Brannan SK, Mahurin RK, Jerabek PA, Brickman JS, Tekell JL, Silva JA, McGinnis S, Glass TG, Martin CC, Fox PT (1997) Cingulate function in depression: A potential predictor of treatment response. Neuroreport 8:1057–1061
38. Mergenthaler E (1998) Cycles of emotion-abstraction patterns: A way of practice oriented process research? Br Psychol Soc – Psychother Sect Newsl 24:16–29
39. Meyer-Lindenberg A, Ziemann U, Hajak G, Cohen L, Faith Berman K (2002) Transitions between dynamical states of differing stability in the human brain. PNAS USA 99:10948–10953
40. Nakao T, Nakagawa A, Yoshiura T, Nakatani E, Nabeyama M, Yoshizato C, Kudoh A, Tada K, Yoshioka K, Kawamoto M, Togao O, Kanba S (2005) Brain activation of patients with obsessive-compulsive disorder during neuropsychological and symptom provocation tasks before and after symptom improvement: A functional magnetic resonance imaging study. Biol Psychiatry 57:901–910
41. Nowak A, Vallacher RR (1998) Dynamical social psychology. Guilford, New York
42. Orlinsky DE, Ronnestad MH, Willutzki U (2004) Fifty years of psychotherapy process-outcome research: Continuity and change. In: Lambert MJ (ed) Bergin and Garfield's handbook of psychotherapy and behavior change. Wiley, New York, pp 307–389
43. Pare D, Curro'Dossi R, Steriade M (1990) Neuronal basis of the parkinsonian resting tremor: A hypothesis and its implications for treatment. Neurosci 35:217–226
44. Penny WD, Stephan KE, Mechelli A, Friston KJ (2004) Modelling functional integration: A comparison of structural equation and dynamic causal models. NeuroImage 23(Suppl 1):264–274
45. Perlitz V, Cotuk B, Haberstock S, Kahn N, Grebe R, Petzold ER, Schmid-Schönbein H (2003) Differentiation of cutaneous haemo- and neurodynamics using multiscale Time-Frequency-Distribution portrays. In: Blazek V, Schultz-Ehrenburg U (eds) Proceedings of the 10th international symposium CNVD 2001 at Aachen. Germany
46. Perlitz V, Cotuk B, Lambertz M, Grebe R, Schiepek G, Petzold ER, Schmid-Schönbein H, Flatten G (2004) Coordination dynamics of circulatory and respiratory rhythms during psychomotor relaxation. Autonom Neurosci 115(1–2):82–93
47. Perlitz V, Cotuk B, Schiepek G, Sen A, Haberstock S, Schmid-Schönbein H, Petzold ER, Flatten G (2004) Synergetik der hypnoiden Relaxation. Psychother Psych Med 54:250–258
48. Perlitz V, Lambertz M, Cotuk B, Grebe R, Vandenhouten R, Flatten G, Petzold ER, Schmid-Schönbein H, Langhorst P (2004) Cardiovascular rhythms in the 0.15 Hz band: Common origin of identical phenomena in man and canine in the reticular formation of the brain stem? Pflügers Arch – Europ J Physiol 448(6):579–592
49. Piaget J (1976) Die Äquilibration der kognitiven Strukturen. Klett-Cotta, Stuttgart
50. Popovych OV, Hauptmann C, Tass PA (2006) Control of neural synchrony by nonlinear delayed feedback. Biol Cybern 95:69–85
51. Rapp PE, Albano ME, Zimmerman ID et al (1994) Phase-randomized surrogates can produce spurious identifications of non-random structure. Phys Lett A 192:27–33
52. Rockstroh B, Watzl H, Kowalik ZJ, Cohen R, Sterr A, Müller M, Elbert T (1997) Dynamical aspects of the EEG in different psychopathological states in an interview situation. A pilot study. Schizophr Res 28:77–85
53. Rosenstein MT, Collins JJ, de Luca CJ (1993) A practical method for calculating Largest Lyapunov Exponents from small data sets. Physica D 65:117–134
54. Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, Cohen JD (2003) The neural basis of economic decision-making in the Ultimatum Game. Science 300:1755–1758
55. Saxena S, Rauch SL (2000) Functional neuroimaging and the neuroanatomy of obsessive-compulsive disorder. Psychiatr Clin North Am 23:563–586
56. Schaub H, Schiepek G (1992) Simulation of psychological processes: Basic issues and an illustration within the etiology of a depressive disorder. In: Tschacher W, Schiepek G, Brunner EJ (eds) Self-organization and clinical psychology. Springer, Berlin, pp 121–149

57. Schiepek G, Schoppek W, Tretter F (1992) Synergetics in psychiatry: Simulation of evolutionary patterns of schizophrenia on the basis of nonlinear difference equations. In: Tschacher W, Schiepek G, Brunner EJ (eds) Self-organization and clinical psychology. Springer, Berlin, pp 163–194
58. Schiepek G, Kowalik ZJ, Schütz A, Köhler M, Richter K, Strunk G, Mühlnickel W, Elbert T (1997) Psychotherapy as a chaotic process I. Coding the client-therapist-interaction by means of sequential plan analysis and the search for chaos: A stationary approach. Psychother Res 7:173–194
59. Schiepek G, Tominschek I, Karch S, Mulert C, Pogarell O (2007) Neurobiologische Korrelate der Zwangsstörungen – Aktuelle Befunde zur funktionellen Bildgebung. Psychother Psych Med 57:379–394
60. Schiepek G, Tominschek I, Karch S, Lutz J, Mulert C, Born C, Pogarell O (2008) A controlled single case study with repeated fMRI measures during the treatment of a patient with obsessive-compulsive disorder: Testing the nonlinear dynamics approach to psychotherapy. World J Biol Psychiatry. doi:10.1080/15622970802311829
61. Schubert C, Schiepek G (2003) Psychoneuroimmunologie und Psychotherapie: Psychosozial induzierte Veränderungen der dynamischen Komplexität von Immunprozessen. In: Schiepek G (ed) Neurobiologie der Psychotherapie. Schattauer, Stuttgart, pp 485–508
62. Schupp HAT, Lutzenberger W, Birbaumer N, Miltner W, Braun C (1994) Neurophysiological differences between perception and imagery. Cog Brain Res 2:77–86
63. Selye HA (1936) Syndrome produced diverse nocuous agents. Nature 138:32
64. Singer W, Gray CM (1995) Visual feature integration and the temporal correlation hypothesis. Ann Rev Neurosci 18:555–586
65. Sinha R, Lovallo WR, Parsons OA (1992) Cardiovascular differentiation of emotions. Psychosom Med 54:422–435
66. Skinner JE, Molnar M, Tomberg C (1994) The point correlation dimension: Performance with nonstationary surrogate data and noise. Int Physiol Behav Sci 29:217–234
67. Stadler M, Kruse P (1990) The self-organization perspective in cognition research. Historical remarks and new experimental approaches. In: Haken H, Stadler M (eds) Synergetics of cognition. Springer, Berlin, pp 32–52
68. Strauss JS (1989) Intermediäre Prozesse in der Schizophrenie: Zu einer neuen dynamisch orientierten Psychiatrie. In: Böker W, Brenner HD (eds) Schizophrenie als systemische Störung. Huber, Bern, pp 39–50
69. Strunk G (2004) Organisierte Komplexität. Mikroprozess-Analysen der Interaktionsdynamik zweier Psychotherapien mit den Methoden der nichtlinearen Zeitreihenanalyse. Dissertation, Universität Bamberg
70. Strunk G, Schiepek G (2006) Systemische Psychologie. Eine Einführung in die komplexen Grundlagen menschlichen Verhaltens. Spektrum Akademischer Verlag, Heidelberg
71. Tallon-Baudry C, Bertrand O (1999) Oscillatory gamma activity in humans and its role in object representation. Trends Cog Sci 3:151–162
72. Tallon-Baudry C, Bertrand O, Wienbruch C, Ross B, Pantev C (1997) Combined EEG and MEG recordings of visual 40 Hz resonses to illusory triangles in human. Neuroreport 8:1103–1107
73. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology (1996) Heart rate variability. Standards of measurement, physiological interpretation, and clinical use. Eur Heart J 17:354–381
74. Tass P, Hauptmann C (2007) Therapeutic modulation of synaptic connectivity with desynchronizing brain stimulation. Int J Psychophysiol 64:53–61
75. Tass PA, Klosterkötter J, Schneider F, Lenartz D, Kouluosakis A, Sturm V (2003) Obsessive-compulsive disorder: Development of demand-controlled deep brain stimulation with methods of stochastic phase resetting. Neuropsychopharmacology 28:S27–S34
76. Tretter F, Scherer J (2006) Schizophrenia, neurobiology, and the methodology of systemic modelling. Pharmacopsychiatry 39(Suppl 1):S26–S35
77. Tschacher W (1997) Prozessgestalten. Hogrefe, Göttingen
78. Tschacher W, Kupper Z (2002) Time series models of symptoms in schizophrenia. Psychiatry Res 113:127–137
79. Tschacher W, Dauwalder JP (1999) Situated cognition, ecological perception, and synergetics: A novel perspective for cognitive psychology? In: Tschacher W, Dauwalder JP (eds) Dynamics, synergerics, autonomous agents. World Scientific, Singapore, pp 83–104
80. Vandenhouten R (1998) Analyse instationärer Zeitreihen komplexer Systeme und Anwendungen in der Physiologie. Shaker, Aachen
81. van Veen V, Carter CC (2002) The anterior cingulate as a conflict monitor: fMRI and ERP studies. Physiol Behav 77:477–482
82. van Veen V, Carter CC (2002) The timing of action-monitoring processes in the anterior cingulate cortex. J Cog Neurosci 14:593–602
83. Villmann T, Liebers C, Bergmann B, Gumz A, Geyer M (2008) Investigation of psycho-physiological interactions between patient and therapist during a psychodynamic therapy and their relation to speech in terms of entropy analysis using a neural network approach. New Ideas in Psychology 26:309–325
84. von Uexküll T, Wesiack W (1996) Wissenschaftstheorie: Ein bio-psycho-soziales Modell. In: Adler RH, Herrmann JM, Köhle K, Schonecke OW, von Uexküll T, Wesiack W (eds) Thure von Uexküll. Psychosomatische Medizin. Urban Schwarzenberg, München, pp 13–52
85. Webber CL, Zbilut JP (1994) Dynamical assessment of physiological systems and states using recurrence plot strategies. J Appl Physiol 76:965–973
86. Yuru Z, Jan KM, Ju KH, Chon KH (2006) Quantifying cardiac sympathetic and parasympathetic nervous activities using principal dynamic modes analysis of heart rate variability. Am J Physiol Heart Circ Physiol 291:H1475–H1483

Self-Organization in Magnetohydrodynamic Turbulence

PIERLUIGI VELTRI, VINCENZO CARBONE,
FABIO LEPRETI, GIUSY NIGRO
Dipartimento di Fisica, Università della Calabria,
Arcavacata di Rende, Italy

Article Outline

Glossary

Alfvén speed The Alfvén speed c_{A} is the propagation speed of Alfvén waves and is given by $c_{\text{A}} = B_0/(4\pi\rho)^{1/2}$, where $\mathbf{B}_0$ is mean magnetic field and ρ the plasma mass density. Alfvèn waves are transverse, incompressible magnetohydrodynamic waves that propagate along $\mathbf{B}_0$ and originate from the tension of magnetic field lines.

Elsässer variables Elsässer variables $\mathbf{z}^{\sigma}$ are defined by $\mathbf{z}^{\sigma} = \mathbf{v} + \sigma\mathbf{B}/(4\pi\rho)^{1/2}$, with $\sigma = \pm 1$, $\mathbf{v}$ the velocity field, $\mathbf{B}$ the magnetic field and ρ the plasma mass density. The equations of incompressible MHD are often written in terms of these variables in order to describe the propagation of Alfvén waves and the non-linear couplings occurring in MHD turbulence.

Magnetohydrodynamics Magnetohydrodynamics (abbreviated, MHD) represents a one-fluid mathematical model which describes plasma dynamics at low frequencies: The main dynamical variables are the velocity of the fluid and the magnetic field. The vector equations for these variables are the fluid momentum conservation and the induction Maxwell equation.

MHD turbulence MHD turbulence is that turbulence which develops inside plasmas at macroscopic level, when viscosity and resistivity are low. Apart from velocity fluctuations which are also present in ordinary fluids, it is characterized also by the presence of magnetic field fluctuations.

Reverse field pinch Reverse field pinch (abbreviated, RFP) are plasma fusion toroidal devices whose conception is based on the idea that non-linear interactions in plasmas spontaneously give rise to magnetic structures where the Laplace force is zero (force free structures).

Shell models Shell models of turbulence are dynamical systems consisting of a set of ordinary differential equations representing a simplified version of the Navier–Stokes or MHD equations in the wavevector space. These models provide the possibility to investigate turbulence at very high Reynolds number regimes at the cost of neglecting information about spatial structures.

Solar corona The solar corona is the region extending from the solar surface up to one million of kilometers in the space, which can be visible to the naked eye during the eclipses. It is constituted mainly by a hydrogen plasma (proton and electrons) at a temperature of about two million degrees. The corona is highly structured by the magnetic field generated at the sun surface.

Solar wind The solar wind is a stream of plasma mainly composed of protons and electrons (hydrogen plasma), which flows out of the sun, due to the fact that plasma pressure associated to the very high coronal temperature overcomes the sun gravity. The flow velocity ranges from 250 km/s in the equatorial plane to about 900 km/s in the polar regions. Solar wind represents an extremely efficient plasma laboratory where the turbulence associated with the supersonic flow can be studied using space experiment data.

Definition of the Subject

Plasma dynamics at low frequency, i. e. at frequencies lower then the ion cyclotron frequency, can be described using a one-fluid model usually called magnetohydrodynamics (MHD), where the main dynamical variables are represented by fluid velocity and magnetic field, which evolve non-linearly being coupled to each other. This description applies both to laboratory plasma devices (tokamaks, reverse field pinch etc.), devoted to realize controlled nuclear fusion, and to space and astrophysical plasmas (Solar corona, Solar wind). Very often, when viscosity and resistivity are sufficiently small, the plasma behavior is characterized by the presence of a developed turbulence. In the last 20 years huge progress in understanding the properties of such turbulence has been realized, both by the use of high resolution computer simulations and by analysis of space and laboratory data. One of the most fascinating results of these studies concerns the evidence of self organization processes which have been shown to be very effective inside this kind of turbulence. Other aspects of turbulence and dynamical complexity in space plasmas are considered in the review by Chang ► Space Plasmas, Dynamical Complexity in in this Encyclopedia.

Introduction

Dynamical systems, whose time evolution is described by non-linear equations, can often give rise to chaotic behavior, i. e. to a behavior characterized by a strong dependence on initial conditions, which, in all practical cases, does not allow to predict the final fate of such systems, or at least limits considerably the possibility to make predictions. Nevertheless, inside these systems, sometimes new complex, coherent structures, without any need of external intervention, can develop. This phenomenon is called self-organization.

Self-organization requires decreasing the entropy inside the coherent organized structures and exporting such entropy to the surroundings. For this reason the coherent structures can be called, following Prigogine, *dissipative structures*, in that their organization is always associated with an increased efficiency of dissipative processes around them.

MHD turbulence, as well as ordinary fluid turbulence, is characterized by a phenomenology, first described some centuries ago around the year 1500 by Leonardo da Vinci (Piumati 1894, fo. 74,v), where energy injected at large scales (injection range) is transferred by non-linear interactions (inertial range) towards smaller and smaller scales where it is finally dissipated (dissipative range). The physical phenomenon by which coherent organized structures are produced through an increased dissipation inside MHD turbulence has been evidenced both at large scale and at small scales. In the latter case it is usually called *intermittency*. We will discuss separately the two cases.

The occurrence of self organization at large scale seems to be directly related to the existence in MHD equations of some quadratic invariants (energy, magnetic helicity, cross helicity), which are conserved inside inviscid flows and are dissipated at very different rates in dissipative flows. Actually the existence of some long-living non-trivial states, towards which dissipative flows are attracted, has been evidenced for a long time. Moreover it has been suggested that these states could be derived from a variational principle, i. e. by minimizing an energy integral subject to some constraints on the other quadratic invariants. Woltjer [58] first explored the astrophysical implications of this possibility. The conjecture by Taylor [51] that an MHD system relaxes towards a state where the energy is minimum, subject to the constraint that magnetic helicity is conserved, allowed one to obtain a quite particular solution, usually called *Taylor's vortex*. This solution was extremely useful to explain the large scale behavior of reverse field pinch (RFP) devices [51]. The same conjecture has also been used to estimate the energy release in coronal structures [25,42] in connection to the problem of coronal heating.

The discovery in the solar wind of quite peculiar *Alfvénic fluctuations*, characterized by a high degree of correlation between velocity and magnetic field [3], which was interpreted by Dobrowolny et al. [20] as the result of a self organization of MHD turbulence, has suggested the formulation [10,38,53] of a different minimum principle: Alfvénic solutions can be obtained by minimizing the energy subject to the condition that cross-helicity is conserved.

In any case the energy minimization is dynamically realized through a *selective decay process*, i. e. a process during which some energy is dissipated into heat and the appearance of the coherent large scale structures is associated with redistribution of entropy to the surroundings.

The self organization processes discussed above, and related to the spontaneous creation of large scale coherent structures during the MHD turbulence dynamics, have their counterpart also at small scales. Actually, the nonlinear energy cascade process in fully developed MHD turbulence is characterized by the formation at the smallest dissipative scale of coherent structures. These structures are strictly related to the intermittency phenomenon; that is, to the breakdown of global self-similarity in the turbulent cascade, which represents a typical signature of nonlinear interactions.

In the following sections we first present the MHD equations (Sect. "MHD Equations and Quadratic Invariants") properties and discuss the characteristics of the self organized structures at large scales (Sect. "Self-Organization at Large Scales"). Then in Sect. "Self-Organization at Small Scales and Intermittent Structures in MHD Turbulence" we show that at small scales coherent dissipative structures are spontaneously formed in a physical process usually called *intermittency*. Finally in Sect. "Future Directions" we discuss the possible future developments and perspectives of such kind of studies.

MHD Equations and Quadratic Invariants

The equations describing the time evolution of an incompressible magnetofluid can be written

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v} = -\frac{1}{\rho}\nabla p + (\nabla\times\mathbf{b})\times\mathbf{b} + \nu\nabla^2\mathbf{v}\,, \quad (1)$$

$$\frac{\partial \mathbf{b}}{\partial t} = \nabla\times(\mathbf{v}\times\mathbf{b}) + \mu\nabla^2\mathbf{b} \quad (2)$$

where $\mathbf{v}$ represents the fluid velocity and p the fluid pressure, $\mathbf{b} = \mathbf{B}/\sqrt{4\pi\rho}$ ($\mathbf{B}$ being the magnetic field and ρ the mass density), ν is the kinematic viscosity and $\mu = (c^2\eta)/(4\pi)$ the magnetic diffusivity (η being the fluid resistivity). The velocity and magnetic field must also satisfy the incompressibility conditions

$$\nabla\cdot\mathbf{v} = 0 \qquad \nabla\cdot\mathbf{b} = 0\,. \quad (3)$$

Quadratic Invariants

It can be easily shown that during their ideal time evolution, i. e. in absence of viscosity and resistivity, these equations conserve the following quadratic invariants: Total energy (kinetic + magnetic) per mass unit

$$E = \frac{1}{2}\int(\mathbf{v}^2 + \mathbf{b}^2)\mathrm{d}V \tag{4}$$

and cross helicity

$$H_\mathrm{c} = \int \mathbf{v}\cdot\mathbf{b}\mathrm{d}V\,. \tag{5}$$

Moreover Woltjer [58] showed that in the ideal 3D MHD, also magnetic helicity

$$H_\mathrm{m} = \int \mathbf{a}\cdot\mathbf{b}\mathrm{d}V \tag{6}$$

($\mathbf{a}$ being the magnetic vector potential defined through $\mathbf{b} = \nabla\times\mathbf{a}$) remains invariant during the evolution of any closed magnetic flux system. In 2D MHD on the contrary, the third conserved quantity is the square of the magnetic field vector potential [24]

$$A = \int \mathbf{a}^2\mathrm{d}V \tag{7}$$

In Eqs. (4), (5), (6) and (7) the integrals are extended to the whole volume of the magnetofluid.

Elsässer's Variables

In the following we will find it useful to rewrite these equations in terms of the Elsässer [21] variables $\mathbf{z}^\sigma$ defined by $\mathbf{z}^\sigma = \mathbf{v} + \sigma\mathbf{b}$ with $\sigma = \pm 1$. Using such variables the equations governing incompressible MHD are recast to the more compact form

$$\frac{\partial\mathbf{z}^\sigma}{\partial t} + (\mathbf{z}^{-\sigma}\cdot\nabla)\mathbf{z}^\sigma = -\frac{1}{\rho}\nabla\left(p + \frac{B^2}{8\pi}\right) + \nu^+\nabla^2\mathbf{z}^\sigma + \nu^-\nabla^2\mathbf{z}^{-\sigma} \tag{8}$$

with $\nu^\sigma = \nu + \sigma\mu$. Obviously also the fields $\mathbf{z}^\sigma$ must satisfy the incompressibility conditions

$$\nabla\cdot\mathbf{z}^\sigma = 0\,. \tag{9}$$

In terms of these variables, the conservation of total energy per mass unit and cross helicity can be written as the conservation of the two pseudo-energies

$$E^\sigma = \frac{1}{2}\int \mathbf{z}^{\sigma^2}\mathrm{d}V\,. \tag{10}$$

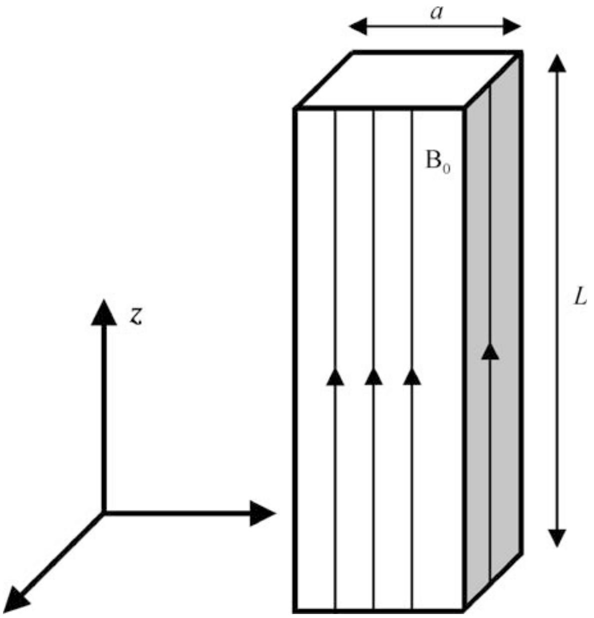
MHD equations display and infinite number of ideal non-quadratic invariants, but quadratic invariants (4), (5) and (6) (or in 2D (7)) play a key role. Actually, let us introduce the Fourier transform of the field $\mathbf{z}^\sigma$ in a cubic periodic box of size a

$$\mathbf{z}^\sigma(\mathbf{r}, t) = \sum_{\mathbf{k}} \mathbf{z}^\sigma(\mathbf{k}, t)\exp(i\mathbf{k}\cdot\mathbf{r})$$

where $\mathbf{k} = 2\pi\mathbf{m}/a$, and $\mathbf{m}$ is vector of integer numbers. In terms of these Fourier transforms (taking into account the incompressibility condition (9) and neglecting dissipative terms), MHD Equations (8) can be written

$$\frac{\partial z_i^\sigma(\mathbf{k}, t)}{\partial t} = \sum_{l,s=1}^{3}\sum_{\mathbf{p},\mathbf{q}} M_{i,l,s}(\mathbf{k}, \mathbf{p}, \mathbf{q}) z_l^\sigma(\mathbf{p}, t) z_s^{-\sigma}(\mathbf{q}, t)\delta_{\mathbf{k},\mathbf{p}+\mathbf{q}} \tag{11}$$

with $M_{i,l,s}(\mathbf{k}, \mathbf{p}, \mathbf{q}) = -i\left(\delta_{il} - \frac{k_i k_l}{k^2}\right)k_s$ and where $\delta_{\mathbf{k},\mathbf{p}+\mathbf{q}}$ is the usual Kronecker symbol. The peculiar form of these equations shows that non-linear terms are characterized by triad interactions among wave vectors $\mathbf{k}$, $\mathbf{p}$ and $\mathbf{q}$ such that $\mathbf{k} = \mathbf{p} + \mathbf{q}$. Quadratic invariants are also called *rugged* invariants, because they are conserved in any single interaction among a triad of wave vector and for this reason



Self-Organization in Magnetohydrodynamic Turbulence, Figure 1
Sketch of the reduced MHD configuration

they survive also when a Galerkin approximation is performed on MHD equations. In other words, even if the r. h. s. of (11) contains an infinite number of non-linear interactions, $E(t)$, $H_c(t)$ and $H_m(t)$ remain invariant also by retaining any finite number of interacting triads in Equation (11).

Reduced Magnetohydrodynamics

In some fusion plasma devices, or in the solar corona, the plasma β parameter (that is the ratio between the kinetic and magnetic pressure) is low $\beta \simeq 10^{-2}$, and MHD equations can be simplified into the so-called Reduced MHD [47,61]. This approximation is valid, for example, for a plasma column with a low aspect ratio $a/L \ll 1$ (a and L being respectively the width and the height of the plasma column), with a strong magnetic field $\mathbf{B}_0$ along the z-direction (Fig. 1).

In this case the plasma dynamical variables reduce to the velocity field $\mathbf{v}_\perp$ and the magnetic field $\mathbf{b}_\perp$ perpendicular to $\mathbf{B}_0$, so that starting from MHD Equs. (8), the following set of RMHD equations [61] is obtained:

$$\frac{\partial \mathbf{z}^\sigma}{\partial t} - \sigma c_A \frac{\partial \mathbf{z}^\sigma}{\partial z} + (\mathbf{z}^{-\sigma} \cdot \nabla_\perp)\mathbf{z}^\sigma = -\frac{1}{\rho}\nabla(p + \frac{B^2}{8\pi}) + \nu^+ \nabla_\perp^2 \mathbf{z}^\sigma + \nu^- \nabla_\perp^2 \mathbf{z}^{-\sigma} \quad (12)$$

$$\nabla_\perp \cdot \mathbf{z}^\sigma = 0 \quad (13)$$

where $\mathbf{z}^\sigma = \mathbf{v}_\perp + \sigma \mathbf{b}_\perp$ (with $\sigma = \pm 1$), $\nabla_\perp$ is the spatial gradient perpendicular to $\mathbf{B}_0$, $c_A = B_0/(4\pi\rho)^{1/2}$ is the Alfvén velocity associated with B_0.

Equation (12) shows that, in this approximation, non-linear interactions are retained only in the directions perpendicular to $\mathbf{B}_0$, while only propagation at the Alfvén speed takes place parallel to $\mathbf{B}_0$.

Let us now Fourier transform the fields with respect to the perpendicular variable $\mathbf{r}_\perp = (x, y)$, in a 2D periodic box of size a

$$\mathbf{v}_\perp(\mathbf{r}, z, t) = \sum_{\mathbf{k}} v(\mathbf{k}, z, t)\mathbf{e}(\mathbf{k}) \exp(i\mathbf{k} \cdot \mathbf{r}) \,, \quad (14)$$

$$\mathbf{b}_\perp(\mathbf{r}, z, t) = \sum_{\mathbf{k}} b(\mathbf{k}, z, t)\mathbf{e}(\mathbf{k}) \exp(i\mathbf{k} \cdot \mathbf{r}) \,, \quad (15)$$

where $\mathbf{k} = 2\pi\mathbf{m}/a$, ($\mathbf{m}$ is a couple of integers) and $\mathbf{e}(\mathbf{k})$ is a unit vector perpendicular to $\mathbf{k}$. After some algebra, it can be shown that in this approximation, for each value of z, the MHD equations reduce to [49]

$$\begin{aligned} \frac{\partial v(\mathbf{k})}{\partial t} &= c_A \frac{\partial b(\mathbf{k})}{\partial z} + \sum_{\mathbf{pq}} c(k, p, q)(p^2 - q^2) \\ &\quad \times \left[v(\mathbf{p})v(\mathbf{q}) - b(\mathbf{p})b(\mathbf{q})\right] \delta_{\mathbf{k},\mathbf{p}+\mathbf{q}} \\ \frac{\partial b(\mathbf{k})}{\partial t} &= c_A \frac{\partial v(\mathbf{k})}{\partial z} + \sum_{\mathbf{pq}} c(k, p, q)k^2 \\ &\quad \times \left[b(\mathbf{p})v(\mathbf{q}) - v(\mathbf{p})b(\mathbf{q})\right] \delta_{\mathbf{k},\mathbf{p}+\mathbf{q}} \end{aligned} \quad (16)$$

(for simplicity we omit the time and z dependence of the Fourier amplitudes) where

$$c(k, p, q) = \frac{p_x q_y - p_y q_x}{2kpq}$$

means that the sum is extended over all wave vectors $\mathbf{p}$ and $\mathbf{q}$ which satisfy the triad-interaction relation $\mathbf{k} = \mathbf{p} + \mathbf{q}$. Let us note that, if we chose a box of size $a = 2\pi$, each wave vector $\mathbf{k}$ turns out to be represented by a couple of integers.

Equations (16) have quadratic invariants namely the total energy

$$E(t) = \sum_{\mathbf{k}} [|v(\mathbf{k}, t)|^2 + |b(\mathbf{k}, t)|^2]$$

and the cross-helicity

$$H_c(t) = \sum_{\mathbf{k}} \mathrm{Re}[v(\mathbf{k}, t)b^*(\mathbf{k}, t)] \,.$$

When the background magnetic field is set to zero $B_0 = 0$, also the mean-square of the vector potential

$$A(t) = \sum_{\mathbf{k}} |b(\mathbf{k}, t)|^2/k^2$$

is conserved. On the contrary, when $B_0 \neq 0$, the last quantity is almost constant, with relative fluctuations of about $\Delta A/A \simeq 10^{-2}$ [49].

This means that 2D MHD has a much wider field of physical applications then those situations where no explicit dependence on the z components exists. Actually 2D MHD represents a good approximation for the low-beta and low aspect ratio plasmas as for example some laboratory devices (tokamaks, RFP, etc.) or some magnetic structures (loops) in the solar corona. In the next section we will take advantage of this result.

Self-Organization at Large Scales

Relaxation Processes in MHD

In MHD, the quadratic invariants (4, 5, 6) play a crucial role in the problem of predicting the final fate of solutions starting from quite general initial conditions. When studying the ideal relaxation, the quadratic invariants remain constant during time evolution, and the problem of predicting the final state reduces to calculating the ensemble–averaged equilibrium spectra of these invariants. Using a standard statistical approach it has been shown by Frisch et al. [23] that these spectra are determined by the ensemble-averaged initial values of invariants. Dissipative relaxation processes in MHD are much more complicated, since the values of the ideal invariants change during the evolution. It seems however that some long-living non-trivial states toward which dissipative flows are attracted exist, and that these states could be derived by minimizing an energy integral subject to some constraints.

Dissipative relaxations have been investigated in connection with measurements in laboratory plasmas, and observations in the solar wind turbulence. Taylor [51], inspired by laboratory plasma experiments, conjectured that an MHD system relaxes towards a state where the energy tends to a minimum, subject to the constraint that magnetic helicity is conserved. The relaxation can then be seen as a *selective decay* between the two invariants. From a mathematical point of view, the solution of the problem can be obtained through a variational principle

$$\delta\left\{\int(\mathbf{v}^2+\mathbf{b}^2)\mathrm{d}V-\lambda\int\mathbf{a}\cdot\mathbf{b}\mathrm{d}V\right\}=0\,,$$

where λ is a Lagrange multiplier. By imposing the conservation of magnetic helicity, we get the so-called force-free solution, characterized by $\mathbf{v}=0$ and

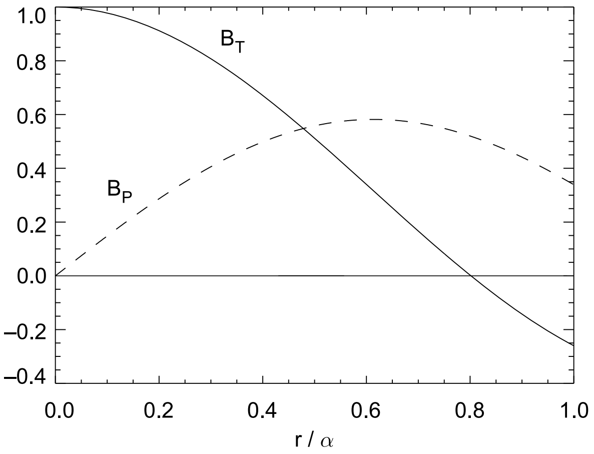
$$\nabla\times\mathbf{b}=\alpha\mathbf{b} \tag{17}$$

(α being a constant), i. e. a solution where the velocity field and the Laplace force are both null. This means that the kinetic energy decays to zero and the magnetic energy occupies the largest scale. Taylor's hypothesis allows one to circumvent in an elegant way the complex plasma relaxation dynamics (involving turbulence and reconnection).

When looking at (17) in cylindrical coordinates, the following solution can be obtained

$$B_T=B_0J_0(\alpha r)\quad B_P=B_0J_1(\alpha r) \tag{18}$$

where B_T and B_P are respectively the axial and poloidal components of the magnetic field (B_0 is a constant and J_0 and J_1 are Bessel functions). This solution, which can display a change in the sign of the magnetic field axial component (Fig. 2), was able to explain the puzzling phenomenon of field reversal spontaneously observed in some laboratory experiments, where no current necessary to produce such reversal was present in the initial state. This theory has also been the basis for the development of RFP devices (Fig. 3), toroidal devices which at a first order approximation can be considered periodic cylinders (Fig. 4). Applications to spheromak and other plasma fusion devices [4,52,59] have been considered more recently.



Self-Organization in Magnetohydrodynamic Turbulence, Figure 2
Axial (*solid line*) and poloidal (*dashed line*) components of magnetic field in a force free cylindrical plasma column as function of the radial coordinate normalized to α^{-1} ($B_0 = 1$)

Belcher and Davis [3] analyzing solar wind data discovered velocity and magnetic field fluctuations were extremely well-correlated (Fig. 5). Dobrowolny et al. [20] suggested that this correlation could be the result of a self organization of MHD turbulence and set up a phenomenological explanation for this self organization, based on the idea that between the two modes of propagation of *Alfvénic* fluctuations in solar wind only that initially dominating survives to the non-linear cascade process towards dissipative scales.

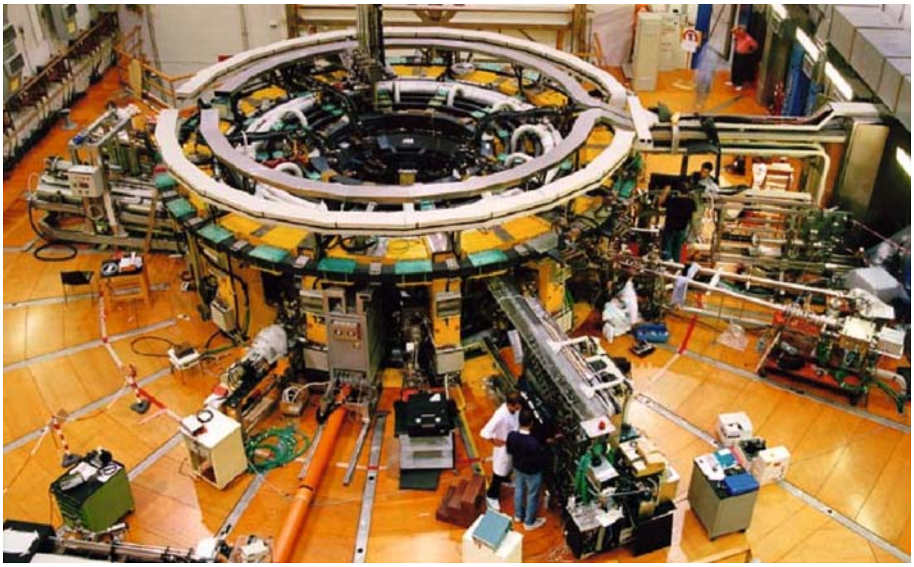
This kind of self organization can also be obtained through a variational principle. Let us impose that energy assumes a minimum value, constrained to the conservation of cross-helicity, we have then to write

$$\delta\left\{\int(\mathbf{v}^2+\mathbf{b}^2)\mathrm{d}V-\lambda\int\mathbf{v}\cdot\mathbf{b}\mathrm{d}V\right\}=0\,,$$

once again λ is a Lagrange multiplier. By imposing the conservation of cross helicity, we finally get the two solutions

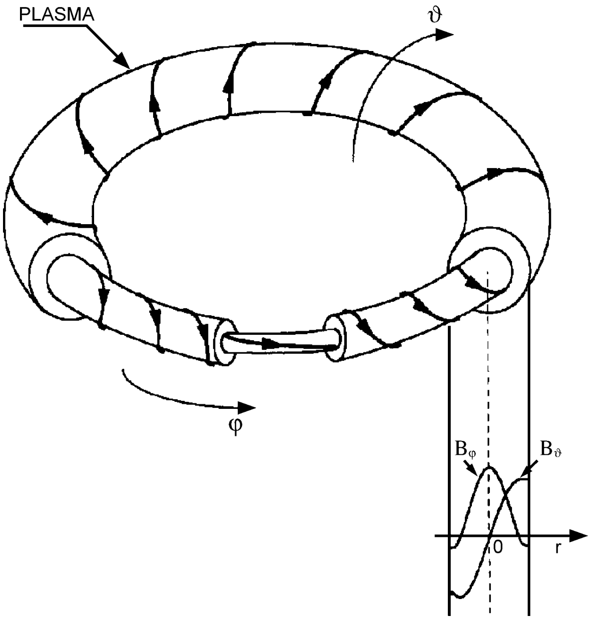
$$\mathbf{v}=\sigma\mathbf{b} \tag{19}$$

with $\sigma=\pm1$. These solutions, which can also be written $\mathbf{z}^\sigma=0$ for either $\sigma=1$ or $\sigma=-1$ in terms of the Elsässer [21] variables, correspond to annihilating non-linear interactions in (1, 8). This phenomenon has been



Self-Organization in Magnetohydrodynamic Turbulence, Figure 3

The RFX: A reverse field pinch plasma device in Padua (Italy) (Consorzio RFX is a research organization promoted by CNR, ENEA, Università di Padova, Acciaierie Venete S.p.A. and INFN, within the framework of the Euratom – ENEA Association)



Self-Organization in Magnetohydrodynamic Turbulence, Figure 4

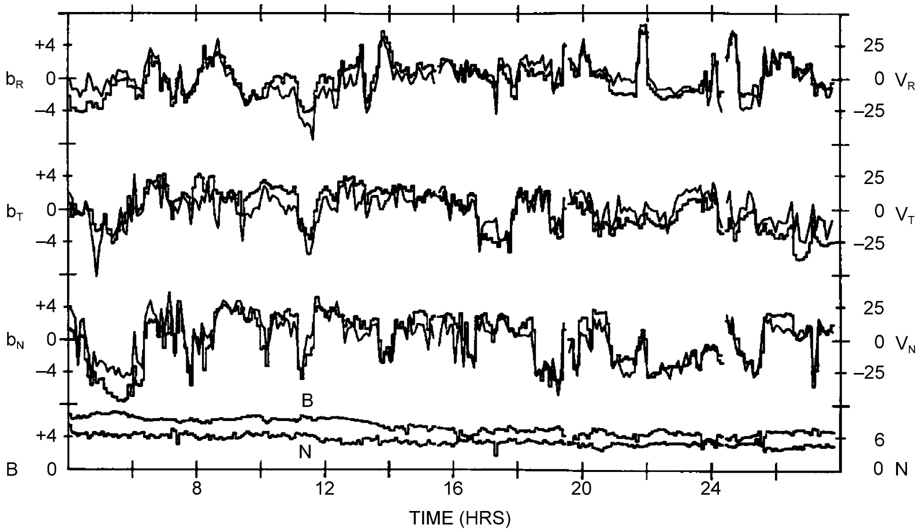
Typical profile of the magnetic field in an RFP device (reprinted from [48])

called *dynamical alignment* in that velocity and magnetic field tend to become aligned to each other.

It has been shown that also in numerical simulations of 2D and 3D MHD equations [17,38,39,46,53] there exists systematic behavior of turbulent flows. In particular, relaxation processes which bring the fluid toward those particular states like *Taylor solutions* or *Alfvénic solutions* have been found to emerge systematically from numerical simulations. In terms of ideal invariants, it has been shown that the Taylor's solution is obtained, when starting with almost zero cross helicity, while the *dynamical alignment* is found when the initial value of the cross helicity is substantially different from zero. In general when both effects are in competition, the situation is much more complicated. A systematic analysis of a wide number of 2D MHD simulations at different resolutions have been performed by Ting et al. [53]. Their results can be summarized by referring to the time behavior of the system projected on the plane (a, h), where $a = A/E$ and $h = 2H_c/E$. In fact, starting from whatever initial conditions, in that plane the system tends to approach the ellipse

$$\frac{1}{a} = 2\left(\frac{k_{\min}}{h}\right)^2 \left[1 \pm \sqrt{1-h^2}\right] , \qquad (20)$$

where $k_{\min}$ represents the minimum wave vector allowed in the simulations. Of course the points $(a, h) = (1/2, \pm 1)$ represents the *attractors* of dynamical alignment solutions, the point $(a, h) = (1, 0)$ represents the attractor of selective decay, and the point $(a, h) = (0, 0)$ represents the attractor when the system behaves like a fluid, with zero magnetic field. These extreme points can be recovered through the minimization of energy subject to the



Self-Organization in Magnetohydrodynamic Turbulence, Figure 5

Velocity (v_R, v_T, and v_N) and magnetic field (b_R, b_T, and b_N) fluctuations measured in solar wind in the RTN reference system (reprinted from [3]). The RTN system has the R unity vector along the radial direction, positive from the sun to the spacecraft, the T unity vector perpendicular to the plane formed by the rotation axis Ω of the sun and the radial direction, i. e., T = Ω × R, and the N component resulting from the vector product N = R × T. Particle density N and magnetic field intensity B are also shown in the lowest part of the plot. Units are 10^{-5} gauss for magnetic fields, km/s for velocities, and cm^{-3} for particle density

conservation of the other invariants. When the magnetic field is zero, the kinetic helicity is invariant and the point $(a, h) = (0, 0)$ can be recovered by allowing that the kinetic energy decays, subject to the constraint that the kinetic helicity remains constant. The entire curve, however, does not represent the locus of the extreme of anything over its entire range of definition.

A Minimal Triad-Interaction Model of Relaxation Processes in 2D MHD

As we have seen, 2D MHD plays a privileged role, in that it describes low-beta and low aspect ratio 3D plasmas. Here we briefly present how the basic non-chaotic triad-interaction works in 2D MHD, and how this represents a minimal basic model for relaxation processes. Choosing only three wave vectors, namely $\mathbf{k}_1 = (1, 1)$, $\mathbf{k}_2 = (2, -1)$, $\mathbf{k}_3 = (3, 0)$, we obtain a set of 12 ordinary differential equations [18] for the complex Fourier modes of both velocity $u_i(t) = u(\mathbf{k}_i, t)/2\sqrt{5}$ and magnetic field $b_i(t) = b(\mathbf{k}_i, t)/2\sqrt{5}$. The system can be further reduced through a projection of equations on a subset of the phase space, that is by considering only real fields. This can be seen by writing the fields in the form $u_j = |u_j| e^{i\alpha_j}$ and $b_j = |b_j| e^{i\beta_j}$, and by defining real fields through $V_j(t) = |u_j| \cos\alpha_j$ and $B_j(t) = |b_j| \cos\beta_j$ (subject to the conditions $\sin\alpha_j = 0$ and $\sin\beta_j = 0$). In this case we found a set of 6 ordinary differential equations for V_j and B_j, namely

$$\begin{aligned}
(\mathrm{d}/\mathrm{d}t + 2\nu)V_1 &= 4(V_3 V_2 - B_3 B_2)\\
(\mathrm{d}/\mathrm{d}t + 5\nu)V_2 &= -7(V_3 V_1 - B_3 B_1)\\
(\mathrm{d}/\mathrm{d}t + 9\nu)V_3 &= 3(V_1 V_2 - B_1 B_2)\\
(\mathrm{d}/\mathrm{d}t + 2\mu)B_1 &= 2(B_3 V_2 - V_3 B_2)\\
(\mathrm{d}/\mathrm{d}t + 5\mu)B_2 &= 5(V_3 B_1 - B_3 V_1)\\
(\mathrm{d}/\mathrm{d}t + 9\mu)B_3 &= 9(V_1 B_2 - B_1 V_2)\,.
\end{aligned}$$

For its simplicity the model represents the basic system to investigate the structure of nonlinear interactions in 2D MHD, and to study the role played by the rugged invariants during the dynamical evolution.

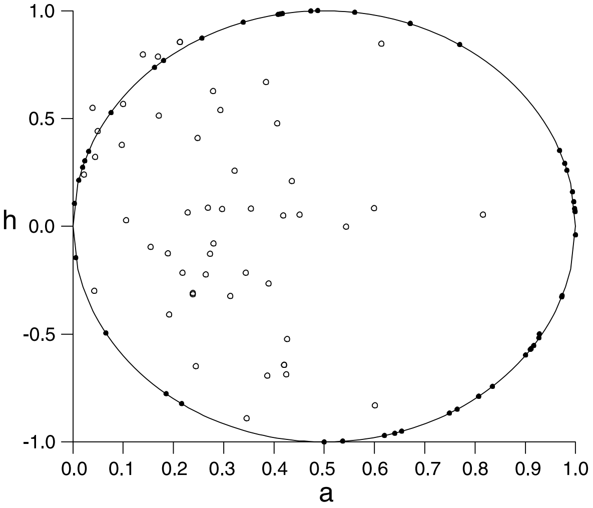
Let us introduce the phase space Ω, of dimension $\mathrm{Dim}[\Omega] = 6$, which can be built up by using as coordinates the Fourier amplitudes. A point $\Psi_i(t) \in \Omega$, defined as $\Psi_i(t) := \{[u_i(t), b_i(t)] \in \Omega\}$ represents the system at a certain time, and this point moves in Ω according to the flow $T_\tau[\Psi_i(t)] = \Psi_i(t + \tau)$ which represents the result of the equation of motion for the system. In absence of dissipative terms the phase space volume is conserved

$$H = \sum_i \frac{\partial}{\partial \Psi_i}\left(\frac{\mathrm{d}\Psi_i}{\mathrm{d}t}\right) = 0\,, \qquad (21)$$

where H represents the rate of change of volumes in phase space.

If we define the ideal flow T_τ^{id} as the flow T_τ obtained when $\nu = \mu = 0$, the phase space volume conservation can be expressed as $T_{\tau\to\infty}^{id}[\Psi_i(t)] = \Psi_i(t+\tau)$. In this case, for a given set of initial values $\Psi_i(0)$, the point moves on a hyper-surface $S \subset \Omega$ defined by the initial value of the invariants. In presence of dissipative terms the quadratic invariants decay, and the rate of change of the volume in the phase space is $H = -16(\nu + \mu) \leq 0$. The condition $H \leq 0$ implies that the dissipative flow pushes the system toward the trivial asymptotic state where all the amplitudes of the fields are zero $T_{\tau\to\infty}^{\text{diss}}[\phi(t)] = 0$.

The triad-interaction model is able to capture the dynamics of the quiescent states observed in MHD. We solved our system by starting from random initial conditions uniformly distributed $\Psi_i(0) \in [-1, 1]$. We used a fourth-order Runge–Kutta scheme, with a time step $\Delta t = 10^{-3}$ and $\nu = \mu = 0.01$. This value for the dissipative coefficients allows the nonlinear interactions to have sufficient time to set up the dynamical behavior. In Fig. 6 we report the curve ε along with two ensembles of points. The first ensemble (white circles) represents the set of points (a, h) obtained with some different initial conditions $\Psi_i(0)$ randomly chosen in the interval $[-1; 1]$. The second ensemble (black circles) represents the set of points (a, h) at time $t = 80$ (in unit of time steps), calculated from the set of fields $\Psi_i(t)$ which are obtained through the time evolution of the set $\Psi_i(0)$, that is $\Psi_i(t) = T_t^{\text{diss}}[\Psi_i(0)]$. As it can be seen that all the initial conditions lead to the final state which belongs to the ellipse ε.



Self-Organization in Magnetohydrodynamic Turbulence, Figure 6

Numerical simulations of the triad-interaction model reported in the plane (a, h). We show two sets of different solutions at two different times. *White circles* refer to a set of 45 initial values $\Psi_i(0)$, *black circles* represent the set of point at a given time $t = 80$ (in unit of time steps) obtained from the time evolution of the above initial values $\Psi_i(t = 80) = T_{t=80}^{\text{diss}}[\Psi_i(0)]$. The solutions represented on that plane are such that, after an initial transient, they fall on the ellipse represented as a *full line*

Since nonlinear interactions in the simple triad-interaction model have the same structure as in the 2D MHD equations, the model is able to capture relaxation properties. These properties are the fixed points and some invariant subspaces. The fixed points of the system can be classified as follows:

a) Two Alfvénic fixed points (say $A^\pm$) characterized by $u_i = \pm b_i$;
b) Three fluid fixed points $F(i)$ $(i = 1, 2, 3)$ where all variables but the velocity V_i are zero;
c) Three magnetic fixed points $M(i)$ $(i = 1, 2, 3)$ where all variables but the magnetic field B_i are zero.

A standard analysis of stability can be performed by linearizing the system around each fixed point. We find that $A^\pm$ are always stable, $M(1)$ is the only stable magnetic fixed point, while $F(1)$ and $F(3)$ are stable. The only stable magnetic fixed point $M(1)$ is such that the energy is localized on the minimum wave vector.

Apart from fixed points the system displays some other interesting properties. Looking at the equations, it can be easily shown that there exists some 3D subspaces of the 6D phase space which remain invariant under the ideal flow operator. Let us denote by I^α the αth invariant subspace, which is then characterized by the fact that if $\Psi_i(0) \in I^\alpha$ then $\Psi_i(t) = T_t^{id}[\Psi_i(0)] \in I^\alpha$ for each $t > 0$. In other words an ideal invariant subspace is a portion of the phase space where the system lies for all times. The most useful way to classify these structures is through the initial values the rugged invariants assume on them, since under the ideal flow they remain constant:

a) A fluid subspace F, characterized by $a = A/E = 0$ and $h = 2H_c/E = 0$. This can be recovered by imposing that the magnetic field is always zero, namely

$$\Psi_i[F] = (V_1, V_2, V_3, 0, 0, 0)$$

is the vector which describes this subspace.

b) Two Alfvénic subspaces $A^\pm$, characterized by $a = A/E = 0$ and $h = 2H_c/E = 1$, which can be recovered by imposing that $V_i = \pm B_i$ for each $i = 1, 2, 3$. These subspaces are also fixed points of the system.

c) Three magnetic subspaces H^A, H^B and H^C characterized by $a \neq 0$ and $h = 0$, i. e. a minimal value for the cross-helicity. These subspaces can be recovered by imposing that the cross-helicity is initially zero over all wave vectors, namely either $V_i = 0$ and $B_i \neq 0$ or vice versa; specifically,

$$\begin{aligned}\Psi_i[H^A] &= (0, 0, V_3, B_1, B_2, 0)\\ \Psi_i[H^B] &= (V_1, 0, 0, 0, B_2, B_3)\\ \Psi_i[H^C] &= (0, V_2, 0, B_1, 0, B_3)\,.\end{aligned}$$

The stability properties of each subspace can be investigated numerically [11]. Simulations start by putting the system on a given subspace, and by adding a small perturbation on the complementary manifold. The motion is not limited to the 3D subspace, and the problem of the stability of a particular I^α consists of determining if the perturbed solution remains close to I^α or goes away from it covering all the allowed 6D phase space. For each subspace we can define two pseudo-energies $E^{(\alpha)}_{\text{int}}(t)$ (built up with the fields which belong to the αth subspace) and $E^{(\alpha)}_{\text{ext}}(t)$ (built up with the fields which do not belong to the αth subspace). The external energy represents the distance $\|\Delta^{(\alpha)}[\Psi_i(t)]\|$ between the point $\Psi_i(t)$ and the αth invariant subspace. Since the total energy $E(t) = E^{(\alpha)}_{\text{ext}}(t) + E^{(\alpha)}_{\text{int}}(t)$ must remain constant under the ideal flow, two situations can arise, namely

1) During the time evolution both $E^{(\alpha)}_{\text{ext}}(t) \simeq E^{(\alpha)}_{\text{ext}}(0)$ and $E^{(\alpha)}_{\text{int}}(t) \simeq E^{(\alpha)}_{\text{int}}(0)$, which means that the solution remains trapped near the invariant subspace. In that case, the subspace is ideally stable.
2) During the time evolution energies become comparable $E^{(\alpha)}_{\text{ext}}(t) \simeq E^{(\alpha)}_{\text{int}}(t)$, which means that the solution is repelled from the invariant subspace. In that case the subspace is ideally unstable.

Numerical simulations show that the fluid subspace is always stable. As far as the magnetic subspaces are concerned, we find that H^A is always stable, while H^B and H^C are always unstable. More information about the behavior of the system near the invariant subspaces can be obtained by considering the characteristic of solutions when the dissipative coefficients are set different from zero. In that case, the energies decay in time, but the rate of decay is different, thus indicating a kind of selective dissipation. In particular, looking at the time evolution of their ratios $R^\alpha(t) = E^{(\alpha)}_{\text{ext}}(t)/E^{(\alpha)}_{\text{int}}(t)$, we find that it decays to zero for the ideally stable subspaces, while it settles to a constant value for the unstable subspaces. This means that ideally stable subspaces, namely the Alfvénic, fluid and magnetic H^A represent a kind of attractor for the system, while unstable subspaces, say H^B and H^C, repel solutions.

When we start numerical solutions near the stable subspaces, the system is attracted to the extreme points of the plane (a, h). When we start near the fluid attractor, the system reach the point $(a, h) = (0, 0)$, and when we start near one of the Alfvénic attractors the system is attracted to $(a, h) = (1/2, \pm 1)$. Finally, when we start near the stable magnetic subspace H^A the system is attracted to the extreme point $(a, h) = (1, 0)$. This last point represents a Taylor regime, say the magnetic field lies on the lowest allowed wave vector. On the contrary, when we start near the unstable magnetic attractors H^B or H^C, the system evolves in an erratic way towards any point (a, h) of the curve (20). Each point (a, h) is made by only one mode with wave vector $k_1 = k_{\text{min}} = \sqrt{2}$, so that if $\eta = V_1/B_1$ we have

$$\begin{aligned}a &= \frac{2}{\left(1+\eta^2\right) k^2_{\text{min}}}\\ h &= \frac{\eta}{1+\eta^2}\,.\end{aligned} \tag{22}$$

This last equation is nothing but the parametric equation of the curve (20).

The behavior we have just described can be recovered also when we start the numerical computation with general initial conditions. In that case, we can associate to each initial condition an invariant subspace according to the rule that the αth subspace is associated with the initial condition $\Psi_i(0)$ when the distance $\|\Delta^{(\alpha)}[\Psi_i(0)]\|$ is the minimum one over α. In other words, we associate an initial condition to the nearest subspace. Numerical simulations show that when the initial conditions are associated to the unstable subspaces H^B and H^C, the final point reached by the system will be any point (a, h) on the curve (20). When the initial conditions are associated with one of the stable subspaces, the final state will be one of the extreme points of the curve (20).

Self-Organization at Small Scales and Intermittent Structures in MHD Turbulence

The self organization processes discussed in the previous sections are related to the spontaneous creation of large scale coherent structures during the MHD turbulence dynamics. However, also at small scales the nonlinear energy cascade process in fully developed MHD turbulence is characterized by the presence of self organization and coherent structures. These structures are strictly related to the intermittency phenomenon, that is, the breakdown of global self-similarity in the turbulent cascade.

The turbulence theory proposed by Kolmogorov in 1941 [30] is based on the conjecture that, within the inertial range, the statistical properties of the fluid motions depend only on the mean energy transfer rate ε and on the scale ℓ, and that the nonlinear energy cascade occurring in this range is a self similar (fractal) process. According to this idea, the velocity field increments $\delta\boldsymbol{v}_\ell(\boldsymbol{r}) = \boldsymbol{v}(\boldsymbol{r}+\boldsymbol{\ell}) - \boldsymbol{v}(\boldsymbol{r})$ scale as $\delta v_\ell \sim \ell^{1/3}$. The self similarity of the turbulent cascade implies that the probability density function (PDF) of field increments $P(\delta v_\ell)$ should be invariant under the scale change

$$P(\delta v_\ell) = \ell^{-1/3} F\left(\frac{\delta v_\ell}{\ell^{1/3}}\right) , \qquad (23)$$

which means that PDFs of normalized velocity increments at different scales collapse onto the same curve. Another consequence of the global self similarity is that the structure functions $S_p(\ell) = \langle \delta v_\ell^p \rangle$, usually defined using the longitudinal field increments, that is, $\delta v_\ell(\boldsymbol{r}) = \left[\boldsymbol{v}(\boldsymbol{r}+\boldsymbol{\ell}) - \boldsymbol{v}(\boldsymbol{r})\right] \cdot \frac{\boldsymbol{\ell}}{\ell}$, should follow the scaling law $S_p(\ell) \sim \ell^{p/3}$.

When considering incompressible MHD turbulence, the scaling properties of the field increments $\delta\boldsymbol{z}^\sigma_\ell(\boldsymbol{r}) = \boldsymbol{z}^\sigma(\boldsymbol{r}+\boldsymbol{\ell}) - \boldsymbol{z}^\sigma(\boldsymbol{r})$, expressed here in terms of Elsässer variables $\boldsymbol{z}^\sigma = \boldsymbol{v} + \sigma\boldsymbol{B}/\sqrt{4\pi\rho}$ ($\boldsymbol{B}$ being the magnetic field and ρ the mass density) can be modified with respect to the fluid case by the so-called Alfvén effect, consisting of the fact that nonlinear interactions take place between eddies of different σ, which propagate in opposite directions at the Alfvén velocity along the large scale magnetic field reducing the efficiency of the non linear energy cascade [20,29,31].

In the fluid-like case a Kolmogorov-like scaling can be expected for the field increments, that is, $\delta z^\sigma_\ell \sim \ell^{1/3}$. On the other hand, when the Alfvén effect is at work the Kraichnan's scaling $\delta z^\sigma_\ell \sim \ell^{1/4}$ is recovered. However, in both cases the turbulent energy cascade is supposed to be a self similar process, resulting in a linear behavior of the structure function exponents, that is, $S_p(\ell) \sim \ell^{p/m}$, where $m = 3$ and $m = 4$ for the fluid-like and Alfvenic case respectively.

Several studies on the statistical properties of fields in fully developed turbulence have shown that the PDFs of field increments are not self similar, that is, their shape change with the scale ℓ. The PDFs at large scales are nearly Gaussian, while, as the scale decreases, the PDFs show sharper and sharper peaks and, correspondingly, higher and higher tails. As far as MHD turbulence is concerned, this behavior has been found both in the solar wind [36,50] and in laboratory plasmas [13]. This departure from self similarity has also been inferred from structure function analysis (see e. g. [8,12,13,35]) which have shown that for both velocity and magnetic field fluctuations the scaling exponents are nonlinear functions of p.

This behavior indicates thus that the turbulent, nonlinear energy cascade is not a fractal (self similar), but rather a multifractal process. In other words, small scale fluctuations much larger than their RMS are present in some spatial positions and the corresponding spatial fluctuations of the energy transfer rate $\varepsilon(\boldsymbol{r})$ play a primary role in the cascade process. These small scale fluctuations have been interpreted as the signature of localized coherent structures spontaneously produced by the nonlinear dynamics. The phenomenon described above is known as *intermittency* in fully developed turbulence.

Another important aspect of intermittency is the occurrence of impulsive bursts of energy dissipation which are observed in different systems characterized by the presence of MHD turbulence, e. g. the magnetic loops of the solar corona [1], and laboratory plasma devices such as reversed field pinches [5] and tokamaks [56]. This phenomenon is often referred to as *temporal intermittency*, to distinguish it from the *spatial intermittency* described above, as it consists in the observation of strong bursts of activity in the time series of a quantity which traces the energy dissipation in the system (e. g. radiation intensity). These bursts are separated from each other by intervals of low activity, denoted as laminar times or waiting times.

While temporal intermittency occurring in turbulent systems has been described, until the end of 90s, in the framework of theoretical paradigms other than turbulence (e. g. Self organized criticality [2]), the idea that spatially and temporal intermittency in fully developed turbulence are both due to the underlying nonlinear dynamics of the energy cascade, giving rise to small scale coherent structures, has now gained considerable standing. In the MHD turbulence context this picture has been supported by a number of results based on the analysis of solar wind and solar corona observations and laboratory plasma experimental data.

The intermittent, coherent events can be viewed as localized zones of fluid where phase correlations exist. These structures, which dominate the statistics of small scales, occur as isolated events. They continuously appear and disappear, apparently in a random fashion, at some locations in the fluid, and they carry most of the flow energy. Among the different techniques which can be used to identify such structures within a turbulent signal, a very effective one is based on the use of wavelet transforms [22]. The wavelet transform of a real square integrable signal $f(x)$ (where x is usually the time or a spatial coordinate) is defined as

$$W(x,r) = C_{\psi}^{-1/2} r^{-1/2} \int_{-\infty}^{\infty} \psi\left(\frac{x'-x}{r}\right) f(x')\mathrm{d}x' \,, \quad (24)$$

where r is a scale dilation, x is a position (time) translation, $\psi(x)$ is the so called mother wavelet function, and C_{ψ} is a normalization constant, which must satisfy the admissibility condition $\int_{-\infty}^{+\infty} |k|^{-1}|\hat{\psi}(k)|^2 \mathrm{d}k < \infty$ where $\hat{\psi}(k)$ is the Fourier transform of $\psi(x)$. The wavelet coefficients $W(x,r)$ give a decomposition of $f(x)$ at the scale r as a function of the position x. For each scale, it is possible to identify the position (or the time occurrence) of strong, intermittent events through a method proposed by Farge [22] based on the so-called Local Intermittency Measure (LIM) $l(x,r)$ defined as

$$l(x,r) = \frac{|W(x,r)|^2}{\langle |W(x,r)|^2 \rangle} \,. \quad (25)$$

The identification method of intermittent structures is based on the idea that large values of $l(x,r)$ represent a signature of large fluctuations with respect to the background level. The wavelet coefficients can thus be classified as "passive" if $l(x,r) < F$ or "intermittent" when $l(x,r) > F$, where F is a threshold which can be chosen by using different suitable criteria [9,41,54].

Another possible method to select intermittent bursts [6] consists of defining bursts as the time intervals during which the condition $f(x) \geq f_{\text{th}}$ is satisfied, where the threshold value f_{th} is calculated through the following iterative method. One starts by defining $f_{\text{th}} = \langle f(t) \rangle + a\sigma$ (a is an arbitrary positive number, usually 2 or 3), where the average and the standard deviation are computed from the whole time series. After excluding the values which exceed f_{th}, the average and the standard deviation are calculated again using the remaining part of the time series. This process is repeated until convergence of f_{th} is reached. This procedure allows one to remove strong events and to evaluate the threshold taking into account only the background contribution.

Once intermittent structures are identified, it becomes possible to study their typical profiles and some relevant statistical properties such as the distributions of event energy and of time intervals between successive bursts. Such investigations allow one to shed light on the physical mechanisms underlying the bursting process. We discuss now the properties of intermittent events observed in three different systems where MHD turbulence occurs, namely the solar wind, reversed field pinches devices, and magnetic loops of the solar corona. Related aspects of intermittency and multifractality are discussed in the review by Chang ▶ Space Plasmas, Dynamical Complexity in in this Encyclopedia.

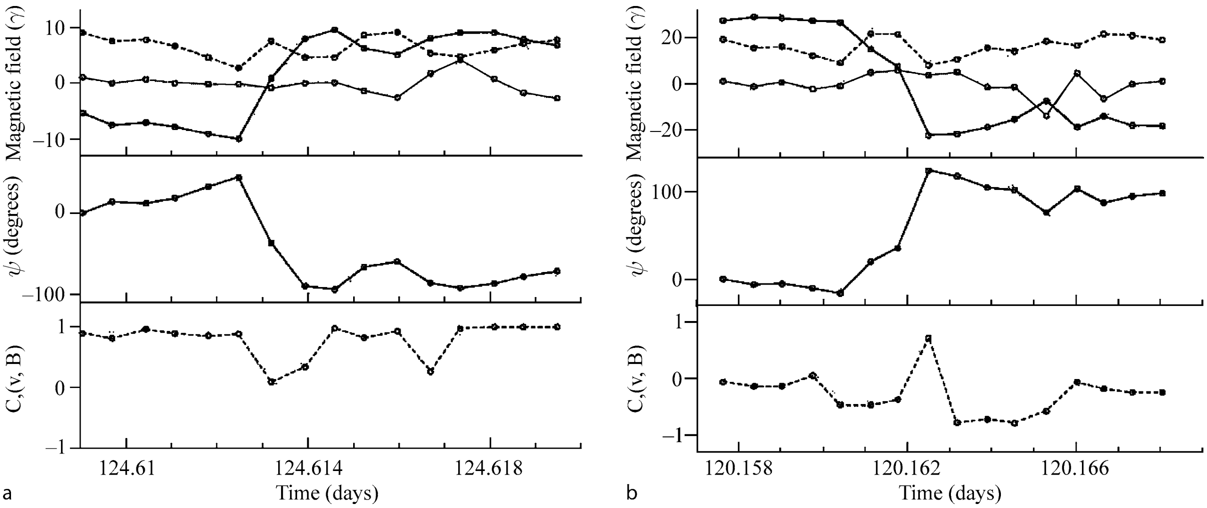
Intermittent Structures in Solar Wind MHD Turbulence

Intermittent events in solar wind turbulence have been studied by Veltri and Mangeney [54] and Veltri et al. [55] by applying the LIM technique, with the Haar wavelet basis, to fluid velocity and magnetic field measurements performed during about 1 year in the space experiment ISEE. In this experiment only two components of the fluid velocity, namely, V_x and V_y, were measured, together with all the three components of the magnetic field. The reference frame used was the standard GSE frame and the sample was formed by data at a time resolution of $T = 1$ min, so that the sampling rate was $\Delta = V_{\text{sw}} T \sim 24.000$ km (V_{sw} is the average solar wind velocity).

The classification of wavelet coefficients allows for an identification and a study of the most intermittent events in solar wind turbulence, which occur in those positions where the amplitude of the wavelet coefficients displays the largest values compared to the average. These events, which occur on time scale of the order of few minutes, exhibit a small number of typical profiles, summarized as follows:

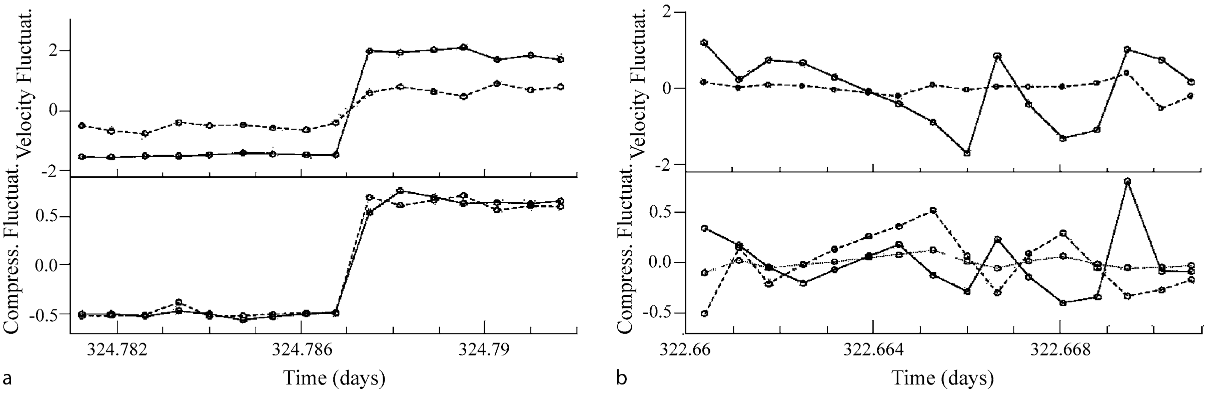
a) "Tangential discontinuities": These structures are almost incompressible, pressure balanced one dimensional current sheets. A minimum variance analysis performed on the magnetic field around the singularity shows (Fig. 7) that the component of the magnetic field which varies most changes sign, and this component is perpendicular to the average magnetic field (the magnetic field component along the third axis being almost zero). The magnetic field rotates then in a plane by an angle which is about 120°–130°. There is one more interesting property: When these structures occur during an Alfvénic period (velocity and magnetic field fluctuations highly correlated), the Alfvénic correlations go from 1 to zero during the traversal of the current sheet (Fig. 7, left panel), when, on the contrary, these structures occur during a period of almost no Alfvénic correlation, the correlation increases to about 1 at the current sheet location (Fig. 7, right panel).

b) "Compressive discontinuities": These structures can be either parallel shocks, mainly observed on the radial component of the velocity field, but clearly seen also on the magnetic field intensity, proton temperature and density (Fig. 8, left panel); or slow mode wavetrains, characterized by a value of $\beta \sim 0$, a constantpressure,



Self-Organization in Magnetohydrodynamic Turbulence, Figure 7

Current sheet intermittent events in solar wind MHD turbulence: The three components of the magnetic field obtained through a minimum variance analysis (*upper panels*); the angle of rotation of the magnetic field in the plane perpendicular to the minimum variance direction (*middle panels*); the coefficient of correlation between velocity and magnetic field fluctuations (*lower panels*)



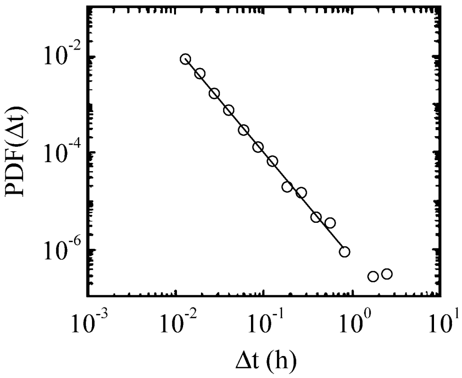
Self-Organization in Magnetohydrodynamic Turbulence, Figure 8

Intermittent events associated with compressive discontinuities in the solar wind. a A parallel shock intermittent event in solar wind, V_x (*full line*) and V_y (*dashed line*) fluctuations normalized to the local average sound velocity (*upper panel*); proton density (*full line*) and sound velocity (*dashed line*) fluctuations normalized to their local average values (*lower panel*). b A slow shock intermittent event in solar wind, velocity fluctuations parallel (*full line*) and perpendicular (*dashed line*) to the local average magnetic field normalized to the local average sound velocity (*upper panel*); proton density (*full line*), sound velocity (*dashed line*) and total pressure (*dotted line*) fluctuations normalized to their local average values (*lower panel*)

anticorrelated density and proton temperature fluctuations and with velocity fluctuations along the average magnetic field (Fig. 8, right panel).

Very interesting statistics can be studied on the time separation Δt (often denoted as *waiting time*) between the occurrence of two consecutive structures. Waiting times of solar wind intermittent structures are distributed according to a power law $P(\Delta t) \sim \Delta t^{-\gamma}$ extended over two decades at least, with an exponent $\gamma \simeq 2.18$ (Fig. 9) [14].

This property is very interesting, because it indicates that the energy cascade process is non-Poissonian. Waiting times occurring between isolated Poissonian events must be distributed according to an exponential function. The power law represents the asymptotic behavior of a Lévy function, which describes self-affine processes and is obtained from the central limit theorem by relaxing the hypothesis that the variance of variables is finite. The power law waiting time PDF is thus a clear evidence that long-range correlations (or in other words "memory") exist in the underlying cascade process.



Self-Organization in Magnetohydrodynamic Turbulence, Figure 9

Probability density functions of waiting times between consecutive intermittent structures in solar wind MHD turbulence identified by applying the LIM technique on magnetic field measurements acquired by the HELIOS II spacecraft. The *solid line* represents a power law with an exponent $\gamma = 2.18$

Intermittent Structures in Laboratory MHD Turbulence

The properties of intermittent events in the magnetic turbulence observed at the edge of a toroidal plasma device in reversed field pinch configuration has been studied by Carbone et al. [14] through the analysis of magnetic field fluctuations in the Reversed Field Pinch experiment RFX [48]. Events have been identified by applying the LIM technique on radial and toroidal magnetic field time series. The magnetic field fluctuations for some of the most intermittent events are shown in Fig. 10 by using the minimum variance reference frame.

One of the two components displays a peak when the other component changes sign. This behavior suggests that the intermittent structures identified in the RFX magnetic field are tangential discontinuities similar to those found in the solar wind, although measurements of all the magnetic field components would be needed to confirm this interpretation.

Also in RFP magnetic turbulence, the waiting time PDF of intermittent structures shows a power law behavior $P(\Delta t) \sim \Delta t^{-\gamma}$ extended over two decades as in the case of solar wind MHD turbulence, with an exponent $\beta \simeq 1.5$ (Fig. 11) [14].

Intermittent Energy Release Events in the Solar Corona

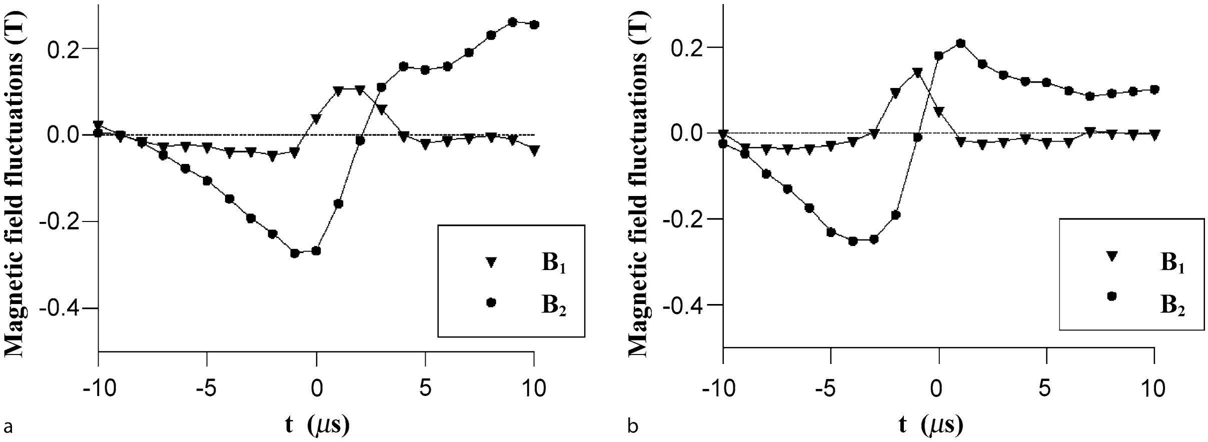
Another manifestation of intermittency phenomena in systems where MHD turbulence is present is the occurrence of impulsive energy release events, the so-called solar flares, which take place in the magnetic loops of the solar corona. Energy is released mainly through accelerated particles and emission of electromagnetic radiation in a wide range of wavelengths, from radio to γ-rays, and can vary between 10^{24} and 10^{33} erg. The smallest events, namely those between 10^{24} and 10^{27}, are usually denoted as nanoflares and microflares.

Parker [43] conjectured that nanoflares are produced by the dissipation of small current sheets, associated with tangential discontinuities, forming as a consequence of the continuous shuffling and intermixing of field footpoints in the photospheric convection. In the Parker's picture, flares of all sizes result from the superposition of small dissipation events (nanoflares) which can trigger energy releases in the neighboring discontinuities, originating a fragmented energy release process which can produce both small and large-scale events, depending on the details of the magnetic configuration.

Starting from the idea that MHD turbulence is most probably a fundamental ingredient of coronal loop (Fig. 12) dynamics, it has been proposed [6] that nanoflares and flares can be identified with dissipation events of small-scale current sheets forming as a consequence of the nonlinear turbulent cascade which occurs inside coronal magnetic structures. The cascade would be driven by the energy input due to photospheric footpoint motions. In this framework, current sheets are coherent intermittent small scale structures of MHD turbulence. The energy injected at large scales by photospheric motions is transferred to small scales through the nonlinear cascade, which goes down to the dissipative scales. The intermittent nature of the energy release process is thus associated with the intermittency of energy dissipation in MHD turbulence.

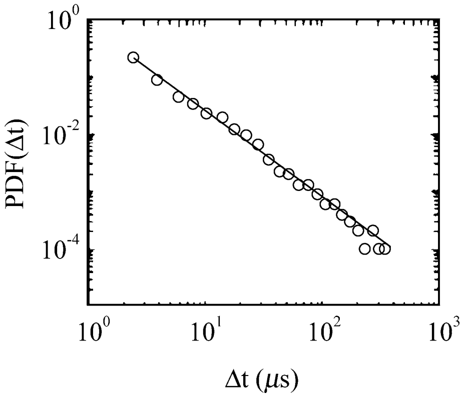
This picture is supported by the statistical properties of solar flares inferred from the analysis of the associated radiation bursts. Nanoflares are detected as extreme ultraviolet (EUV) brightenings, while larger flares are observed through several types of emission (radio, microwaves, $H\alpha$, UV, X-rays) and most often studied analyzing Soft X-ray (SXR) and Hard X-ray (HXR) flare bursts. Probability distributions of the relevant quantities (peak flux, total energy, duration) characterizing the flare bursts have been found to be well represented by power laws $P(x) = Ax^{-\alpha}$ (see Fig. 13 for an example).

Using various SXR and HXR observations it has been found that, for energy and peak flux distributions, $\alpha \simeq 1.6$–2 (see e. g. [15,16,34]). For the EUV nanoflare brightenings the power law exponent is much more uncertain, $\alpha = 1.3$–2.8 (see e. g. [32,44]). The waiting time distribution (WTD) has been studied both for HXR and SXR bursts [6,45,57]. SXR observations acquired by the GOES



Self-Organization in Magnetohydrodynamic Turbulence, Figure 10

Two components of magnetic field fluctuations (denoted as B_1 and B_2), in the minimum variance reference frame, of two intermittent events observed in the RFX machine



Self-Organization in Magnetohydrodynamic Turbulence, Figure 11

Probability density functions of waiting times between consecutive intermittent structures in RFP magnetic turbulence identified by applying the LIM technique on magnetic field measurements acquired at the RFX machine. The *solid line* represents a power law with an exponent $\gamma = 1.5$

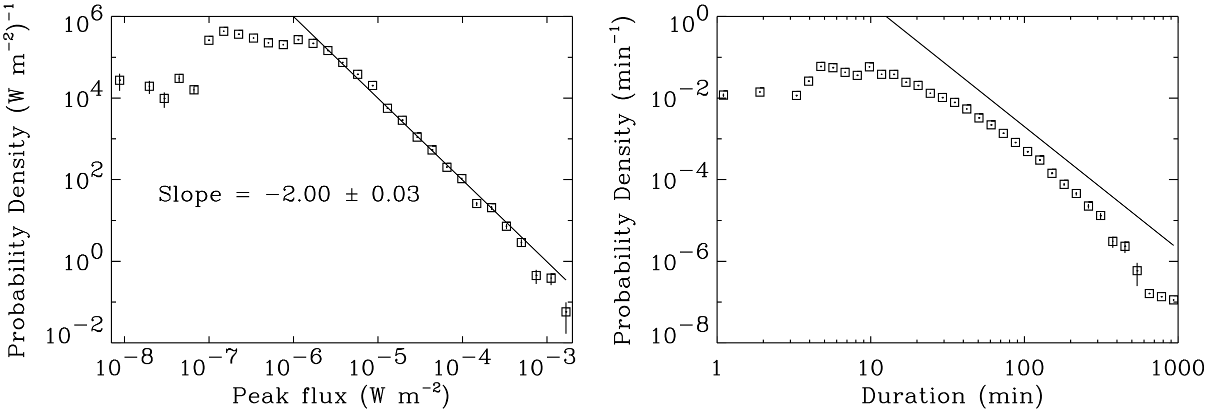
Self-Organization in Magnetohydrodynamic Turbulence, Figure 12

An image of hot coronal loops which span 30 or more times the diameter of planet Earth obtained from Transition Region and Coronal Explore (TRACE) satellite. The image was taken in the Fe IX 171 Å emission line. (Transition Region and Coronal Explorer, TRACE, is a mission of the Stanford–Lockheed Institute for Space Research, and part of the NASA Small Explorer program)

satellites have the advantage to provide a long sequence of bursts (from 1975 to today) with few gaps and allow, thus, to analyze the WTD with a much better statistical accuracy than HXR data. The WTD of GOES SXR flares has been shown to display a clear power law tail $P(\Delta t) \sim \Delta t^{-\gamma}$ in the range $5\,\mathrm{h} \lesssim \Delta t \lesssim 100\,\mathrm{h}$, with an exponent $\gamma \simeq 2.4$ (Fig. 14) [6].

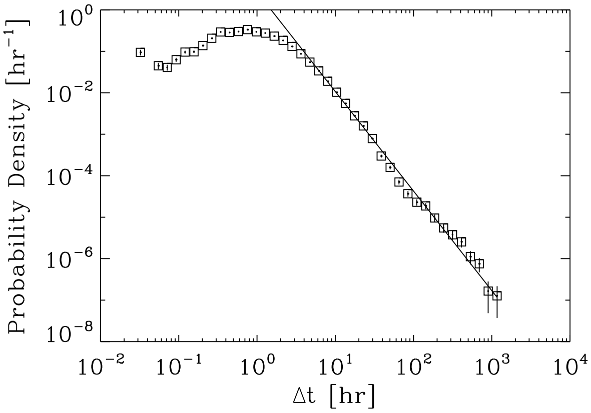
The power law behavior of the solar flare WTD represents an indication of the existence of correlations between successive bursts [6,33]. These correlations can be related to the nonlinear dynamics of the MHD turbulent energy cascade which generates intermittent bursts of chaoticity at small scales.

From the theoretical point of view, describing in a proper way the intermittency of the turbulent energy



Self-Organization in Magnetohydrodynamic Turbulence, Figure 13

Probability density functions of peak flux (*left panel*) and duration (*right panel*) of Soft *X*-ray solar flare bursts detected by the Geostationary Operational Environmental Satellites (GOES). The *solid line* in the *left panel* represents a power law with an exponent $\alpha = 2$ obtained from a least squares fit. The *solid line* in the *right panel*, shown as a comparison, represents a power law with an exponent $\alpha = 3$



Self-Organization in Magnetohydrodynamic Turbulence, Figure 14

Probability density function of waiting times for SXR solar flare bursts detected by the GOES satellites in the interval from September 1975 to December 2001. The solid line represents a power law with an exponent $\alpha = 2.4$ as found by Boffetta et al. [6]

dissipation process is one of the basic ingredients for the study of intermittent events occurring in the solar corona, as well as in other astrophysical systems such as the solar wind and accretion disks. These astrophysical plasmas are characterized by huge Reynolds numbers. The number N_{df} of degrees of freedom, i.e the number of grid points necessary to resolve the entire inertial range of turbulence in a direct numerical simulation of Navier–Stokes or MHD equations, grows with the Reynolds number Re as $N_{\mathrm{df}} \sim \mathrm{Re}^{9/4}$. Therefore performing direct numerical simulations at Reynolds number regimes of interest for space plasmas is out of the present computational possibilities. The so called shell models, a class of dynamical deterministic models of turbulence in which N_{df} grows logarithmically with Re, can represent an extremely helpful tool for the modeling of such physical systems, as these models are able to simulate the turbulent cascade and the related intermittency of the energy dissipation process in Reynolds number regimes which are not far from the real ones (at least with respect to direct numerical simulations). The statistical properties of intermittent events in MHD turbulence can thus be effectively investigated through the use of shell models [6,14,15].

Shell models [7] are dynamical systems designed in order to represent in a simplified way the spectral Navier–Stokes or MHD equations for turbulence. They were originally proposed by Obukhov [40], Desnyansky and Novikov [19], and Gledzer [26] in hydrodynamic turbulence.

In order to build up the evolution equations for a shell model, the wavevector space is divided into a discrete number of shells of radius $k_n = \rho^n k_0$, with $\rho > 1$, $n = 1, 2, \ldots, N$. Each shell is assigned a scalar dynamic variable (real or complex) $v_n(t)$ which represents the averaged effect of velocity modes with wavenumber between k_n and k_{n+1}. $v_n(t)$ can also be regarded as the velocity increments $|v(x+\ell) - v(x)|$ on an eddy of scale $\ell \sim k_n^{-1}$. For MHD shell models, besides $v_n(t)$, another variable $b_n(t)$, representing magnetic field modes between k_n and k_{n+1} (or, alternatively, magnetic field increments $|b(x+\ell) - b(x)|$), is assigned to each shell. The nonlinear terms of the equations are quadratic combinations of the dynamic variables and are written under the assumption that the interactions among shells are local in k-space. This means that only nearest and, at most, next-nearest shells are introduced in the non-linear terms. The coupling coefficients are found by imposing conservation of the ideal invariants of Navier–Stokes or MHD equations.

The main advantage of shell models is that they can be studied through numerical simulations at very high Reynolds numbers. Moreover, they provide the possibility to obtain long time series with a fairly small computational effort, allowing thus a robust investigation of the statistical properties of intermittent events The weak point is that they are scalar models, that is, any information about spatial structures, such as vortices, sheets, and filaments, is lost.

Among the various shell models which have been proposed in the literature, we consider here the so called GOY (Gledzer–Ohkitani–Yamada) model, which has been used by several authors both in hydrodynamics and MHD. In the MHD context, the advantage of the GOY model with respect to previous shell models, where only the conservation of two quadratic invariants, i. e. total energy and cross helicity, can be imposed, is that it allows to conserve also magnetic helicity. The GOY model involves nearest and next-nearest interactions and was originally introduced in the framework of hydrodynamic turbulence by Gledzer [26], using real shell variables. The model was extended by Yamada and Ohkitani [60] to the case of complex variables. The MHD generalization of the GOY model has been considered by several authors (see [6,27,28] and references therein). The evolution equations for the dynamic variables $v_n(t)$ and $b_n(t)$ in the GOY shell model can be written as [6]

$$\begin{aligned}\frac{\mathrm{d}v_n}{\mathrm{d}t} = -\nu k_n^2 v_n + f_n + ik_n \Big\{ &(v_{n+1}v_{n+2} - b_{n+1}b_{n+2}) \\ &- \frac{1}{4}(v_{n-1}v_{n+1} - b_{n-1}b_{n+1}) \\ &- \frac{1}{8}(v_{n-2}v_{n-1} - b_{n-2}b_{n-1}) \Big\}^* , \quad (26)\end{aligned}$$

$$\begin{aligned}\frac{\mathrm{d}b_n}{\mathrm{d}t} = -\mu k_n^2 b_n + g_n + ik_n \frac{1}{6} \Big\{ &(v_{n+1}b_{n+2} - b_{n+1}v_{n+2}) \\ &+ (v_{n-1}b_{n+1} - b_{n-1}v_{n+1}) \\ &+ (v_{n-2}b_{n-1} - b_{n-2}v_{n-1}) \Big\}^* . \quad (27)\end{aligned}$$

f_n and g_n represent external forcing terms, usually acting on low wavenumber shells, while ν and μ appearing in the linear, dissipative terms, are the kinematic viscosity and the resistivity respectively, viscous and resistive terms provide a mechanism for energy dissipation at small scales. In turbulent space plasmas such as solar corona and solar wind dissipation mechanisms are different, involving kinetic plasma processes. Nevertheless, the properties of low frequency MHD turbulent cascade in the inertial range and of the related intermittency do not depend, in general, on the details of the dissipation mechanism, hence the use of viscous and resistive dissipation in the standard incompressible form does not represent a serious limitation in this context.

The GOY model described here has been designed to investigate 3D MHD turbulence. This means that it satisfies the conservation of the ideal 3D MHD quadratic invariants (4), (5) and (6). In terms of shell variables, the invariants read [27,28]

$$E = \frac{1}{2}\sum_{n=1}^{N}(|v_n|^2 + |b_n|^2) , \quad (28)$$

$$H_\mathrm{c} = \frac{1}{4}\sum_{n=1}^{N}\Re(v_n b_n^*) , \quad (29)$$

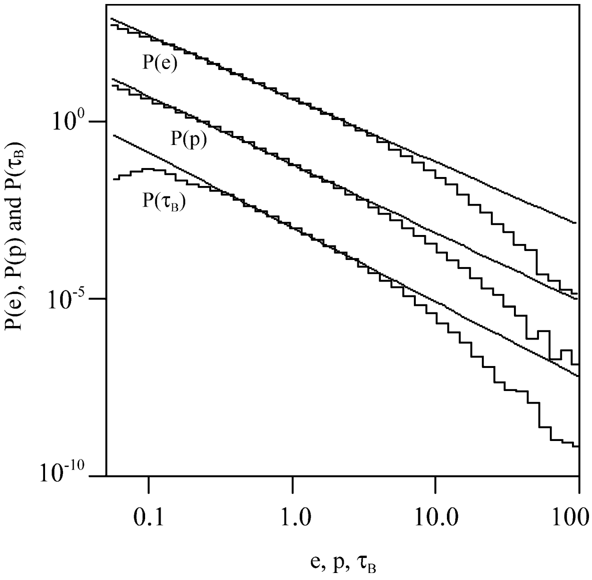
$$H_\mathrm{m} = \sum_{n=1}^{N}(-1)^n \frac{|b_n|^2}{k_n} . \quad (30)$$

The energy dissipation rate in the GOY MHD shell model can be defined as [6]

$$\varepsilon(t) = \nu \sum_{n=1}^{N} k_n^2 |v_n|^2 + \mu \sum_{n=1}^{N} k_n^2 |b_n|^2 . \quad (31)$$

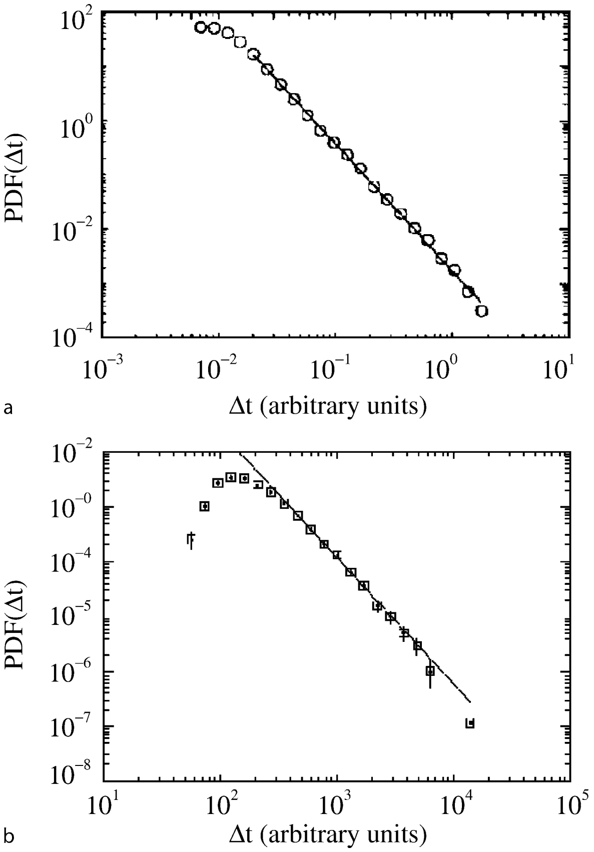
A statistical analysis of intermittent events in the GOY MHD shell model shows that the PDFs of event energies, duration and waiting times display power law tails (Figs. 15 and 16) [6,14]. This behavior is consistent with the statistics of intermittent events observed in the solar wind, reversed field pinch devices and solar corona.

In closing this section, it can thus be remarked that the study of intermittent events in MHD turbulence, mainly based on wavelet transforms and PDFs of various event parameters, has provided a quite detailed picture of the coherent structures produced by the self-organization of the MHD turbulence cascade at small scales. The analysis of data from solar wind and RFP devices has demonstrated that the most intermittent structures in incompressible MHD are one dimensional current sheets associated with tangential discontinuities. The statistical analysis of intermittent events occurring in solar wind, laboratory plasma and magnetic structures of the solar corona has shown that these events are characterized by power law distributions of event energy and waiting times, which are naturally reproduced by MHD turbulence models. The power law behavior of the waiting time distribution indicates the



Self-Organization in Magnetohydrodynamic Turbulence, Figure 15

Total energy distribution $P(e)$, peak energy distribution $P(p)$ and burst duration distribution $P(\tau_B)$ for the GOY MHD shell model. The values of $P(e)$ and $P(\tau_B)$ are offset by a factor 10^2 and 10^{-2} respectively. The straight lines represent power law fits, with exponents $\alpha_e = 1.8$, $\alpha_p = 2.05$ and $\alpha_{\tau_B} = 2.2$ respectively



Self-Organization in Magnetohydrodynamic Turbulence, Figure 16

Probability density function of waiting times between successive intermittent events in the GOY MHD shell model. a Waiting time PDF for intermittent events identified in the time series of the magnetic field variables at the largest wavenumber shell. The *solid line* represents a power law with exponent $\beta = 2.13$. b Waiting time PDF for intermittent events identified in the time series of the energy dissipation rate ε (Eq. 31). The solid line represents a power law with exponent $\beta = 2.33$. In both cases, intermittent events were identified through the iterative procedure proposed in [6] and described in the text, using the threshold $\langle\cdot\rangle + 2\sigma$

presence of long range time correlations which can be attributed to the non-linear dynamics of the turbulent energy cascade.

Future Directions

In the previous sections we have discussed how coherent dissipative structures are spontaneously formed by non-linear interactions both in laboratory and in natural plasmas. A further understanding of these processes requires an increase in the performance of computers, which will allow one to simulate 3D MHD turbulence at Reynolds numbers larger than 10^3, which represents the actual limit. On the other hand, an improved effort in the analysis of solar wind space data and laboratory plasma devices measurements should also be necessary. An increased knowledge of these phenomena is crucial in order to try to control the transport processes in fusion plasmas and then to improve the plasma confinement, since the presence of coherent structures inside turbulence greatly affect the plasma diffusion process. Moreover a better comprehension of the self organization in MHD turbulence can greatly help in explaining some basic phenomena in solar corona (flares and coronal heating), which occur in a region where in situ observations cannot be performed.

Bibliography

Primary Literature

1. Aschwanden MJ (2004) Physics of the solar corona: An introduction. Springer, Berlin
2. Bak P, Tang C, Wiesenfeld K (1987) Self-organized criticality: An explanation of 1/*f* noise. Phys Rev Lett 59:381–384
3. Belcher JW, Davis L (1971) Large-amplitude Alfvèn waves in the interplanetary medium-II. J Geophys Res 76:3534–3563
4. Bellan PM (2000) Spheromaks: A practical application of magnetohydrodynamic dynamos and plasma self-organization. Imperial College Press, London

5. Biskamp D (1993) Non-linear magnetohydrodynamics. Cambridge University Press, Cambridge
6. Boffetta G, Carbone V, Giuliani P, Veltri P, Vulpiani A (1999) Power laws in solar flares: Self-organized criticality or turbulence? Phys Rev Lett 83:4662–4665
7. Bohr T, Jensen MH, Paladin G, Vulpiani A (1998) Dynamical systems approach to turbulence. Cambridge University Press, Cambridge
8. Burlaga LF (1991) Multifractal structure of the interplanetary magnetic field Voyager 2 observations near 25 AU, 1987 1988. Geophys Res Lett 18:69–72
9. Camussi R, Guj G (1997) Orthonormal wavelet decomposition of turbulent flows: Intermittency and coherent structures. J Fluid Mech 348:177–199
10. Carbone V, Veltri P (1987) A simplified cascade model for M.H.D. turbulence. Astron Astrophys 188:239–250
11. Carbone V, Veltri P (1992) Relaxation processes in MHD: A triad-interaction model. Astron Astrophys 259:359–372
12. Carbone V (1994) Scaling exponents of the velocity structure functions in the interplanetary medium. Ann Geophys 12:585–590
13. Carbone V, Sorriso-Valvo L, Martines E, Antoni V, Veltri P (2000) Intermittency and turbulence in a magnetically confined fusion plasma. Phys Rev E 62:49–56
14. Carbone V, Cavazzana R, Antoni V, Sorriso-Valvo L, Spada E, Regnoli G, Giuliani P, Vianello N, Lepreti F, Bruno R, Martines E, Veltri P (2002) To what extent can dynamical models describe statistical features of turbulent flows? Europhys Lett 58:349–355
15. Carbone V, Lepreti F, Sorriso-Valvo L, Veltri P, Antoni V, Bruno R (2004) Scaling laws in plasma turbulence. Riv Nuovo Cim 27(8-9):1–108
16. Crosby NB, Aschwanden MJ, Dennis BR (1993) Frequency distributions and correlations of solar *X*-ray flare parameters. Solar Phys 143:275–299
17. Dahlburg JP, Montgomery D, Doolen GD, Turner R (1986) Turbulent relaxation to a force-free field-reversed state. Phys Rev Lett 57:428–431
18. De Bartolo R, Carbone V (2006) The role of the basic three-modes interaction during the free decay of magnetohydrodynamic turbulence. Europhys Lett 73:547–552
19. Desnyansky VN, Novikov EA (1974) Evolution of turbulence spectra to self-similar regime. Atmos Oceanic Phys 10:127–136
20. Dobrowolny M, Mangeney A, Veltri P (1980) Fully developed anisotropic hydromagnetic turbulence in interplanetary space. Phys Rev Lett 45:144–147
21. Elsässer WM (1950) The hydromagnetic equations. Phys Rev 79:183–183
22. Farge M (1992) Wavelet transforms and their applications to turbulence. Ann Rev Fluid Mech 24:395–457
23. Frisch U, Pouquet A, Leorat J, Mazure A (1975) Possibility of an inverse cascade of magnetic helicity in magnetohydrodynamic turbulence. J Fluid Mech 68:769–778
24. Fyfe D, Montgomery D (1976) High-beta turbulence in two-dimensional magnetohydrodynamics. J Plasma Phys 16: 181–191
25. Heyvaerts J, Priest ER (1984) Coronal heating by reconnection in DC current systems – A theory based on Taylor's hypothesis. Astronom Astrophys 137:63–78
26. Gledzer EB (1973) System of hydrodynamic type admitting two quadratic integrals of motion. Sov Phys Dokl 18:216–217
27. Giuliani P, Carbone V (1998) A note on shell models for MHD turbulence. Europhys Lett 43:527–532
28. Giuliani P (1999) Non-linear MHD waves and turbulence, edited by Passot T, Sulem PL. Lect Notes Phys 536:331–345
29. Iroshnikov PS (1963) Turbulence of a conducting fluid in a strong magnetic field. Astron Zhur 40:742–750 (in Russian, translated in Sov Astron 7:566–571)
30. Kolmogorov AN (1941) The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers. Dokl Akad Nauk SSSR 30:299–303 (in Russian translated in Proc Royal Soc London 434:9–13)
31. Kraichnan RH (1965) Inertial range spectrum of hydromagnetic turbulence. Phys Fluids 8:1385–1387
32. Krucker S, Benz AO (1998) Energy Distribution of Heating Processes in the Quiet Solar Corona. Astrophys J 501:L213–L216
33. Lepreti F, Carbone V, Veltri P (2001) Solar flare waiting time distribution: Varying rate Poisson or Lévy function? Astrophys J 555:L133–136
34. Lin RP, Schwartz RA, Kane SR, Pelling RM, Hurley KC (1984) Solar hard *X*-ray microflares. Astrophys J 283:421–425
35. Marsch E, Liu S (1993) Structure functions and intermittency of velocity fluctuations in the inner solar wind. Ann Geophys 11:227–238
36. Marsch E, Tu CY (1994) Non-Gaussian probability distributions of solar wind fluctuations. Ann Geophys 12:1127–1138
37. Tu CY, Marsch E (1995) MHD structures, waves and turbulence in the solar wind: Observations and theories. Space Sci Rev 73:1–210
38. Matthaeus WH, Montgomery D (1980) Selective decay hypothesis at high mechanical and magnetic Reynolds numbers. Ann NY Acad Sci 357:203–222
39. Matthaeus WH, Goldstein ML, Montgomery D (1982) Turbulent generation of outward-traveling interplanetary alfvnic fluctuations. Phys Rev Lett 51:1484–1487
40. Obukhov AM (1971) Some general characteristic equations of the dynamics of the atmosphere. Atmos Oceanic Phys 7:41–45
41. Onorato M, Camussi R, Iuso G (2000) Small scale intermittency and bursting in a turbulent channel flow. Phys Rev E 61:1447–1454
42. Parker EN (1983) Magnetic neutral sheets in evolving fields. II. Formation of the solar corona. Astrophys J 264:642–647
43. Parker EN (1988) Nanoflares and the solar *X*-ray corona. Astrophys J 330:474–479
44. Parnell CE, Jupp PE (2000) Statistical Analysis of the Energy Distribution of Nanoflares in the Quiet Sun. Astrophys J 529:554–569
45. Pearce G, Rowe A, Yeung J (1993) A statistical analysis of HRD *X*-ray solar flares. Astrophys Space Sci 208:99–111
46. Pouquet A, Meneguzzi M, Frish U (1986) Growth of correlations in magnetohydrodynamic turbulence. Phys Rev A 33:4266–4276
47. Rosenbluth MN, Monticello DA, Strauss HR (1976) Numerical studies of non-linear evolution of kink modes in tokamaks. Phys Fluids 19:1987–1996
48. Rostagni G (1995) RFX: An expected step in RFP research. Fusion Eng Design 25:301–313
49. Servidio S, Carbone V (2005) Non-linear dynamics of inviscid reduced MHD plasmas: The appearance of quasi-single-helicity states. Phys Rev Lett 95:045001
50. Sorriso-Valvo L, Carbone V, Veltri P, Consolini G, Bruno R (1999) Intermittency in the solar wind turbulence through proba-

bility distribution functions of fluctuations. Geophys Res Lett 26:1801–1804

51. Taylor JB (1974) Relaxation of toroidal plasma and generation of reverse magnetic fields. Phys Rev Lett 33:1139–1141
52. Taylor J (1986) Relaxation and magnetic reconnection in plasmas. Rev Mod Phys 58:741–763
53. Ting AC, Matthaeus WH, Montgomery D (1986) Turbulent relaxation processes in magnetohydrodynamics. Phys Fluids 29:3261–3274
54. Veltri P, Mangeney A (1999) Scaling laws and intermittent structures in solar wind MHD turbulence. In: Habbal SR, Esser R, Hollweg JV, Isenberg PA (eds) Solar wind nine, Proceedings of the ninth international solar wind conference, AIP Conference Proceedings vol 471 pp 543–546
55. Veltri P, Nigro G, Malara F, Carbone V, Mangeney A (2005) Intermittency in MHD turbulence and coronal nanoflares modelling. Non-lin Proc Geophys 12:245–255
56. Wesson J (2003) Tokamaks. Oxford Science, Oxford
57. Wheatland MS, Sturrock PA, McTiernan JM (1998) The waiting-time distribution of solar flare hard *X*-Ray bursts. Astrophys J 509:448–455
58. Woltjer L (1958) A theorem on force-free magnetic fields. Proc Natl Acad Sci 44:489–491
59. Yamada M (1999) Study of magnetic helicity and relaxation phenomena in laboratory plasmas. AGU Geophys Monogr 111:129–140
60. Yamada M, Ohkitani K (1987) Lyapunov spectrum of a chaotic model of three dimensional turbulence. J Phys Soc Japan 56:4210–4213
61. Zank GP, Matthaeus WH (1992) The equations of reduced magnetohydrodynamics. J Plasma Phys 48:85–100.

Books and Reviews

Bruno R, Carbone V (2005) The solar wind as a turbulence laboratory. Living Rev Solar Phys. 2: http://www.livingreviews.org/lrsp-2005-4

Biskamp D (2003) Magnetohydrodynamic turbulence. Cambridge University Press, Cambridge

Frisch U (1995) Turbulence: The legacy of A. N. Kolmogorov. Cambridge University Press, Cambridge

Moffat HK (1978) Magnetic field generation in electrically conducting fluids. Cambridge University Press, Cambridge

Self-organized Criticality and Cellular Automata

MICHAEL CREUTZ
Physics Department,
Brookhaven National Laboratory, Upton, USA

Article Outline

Glossary

Abelian group A mathematical group wherein all the elements commute.

Avalanche A possibly large disturbance induced in a system by a small perturbation.

Cellular automaton This refers to the dynamics of a collection of cells each of which can be in a finite set of states. The evolution is discrete, with the state of a cell at the next time step being dependent only on its previous state and that of its neighbors.

Chaos The tendency of a system of a few degrees of freedom to exhibit highly erratic behavior characterized by an infinite range of time scales.

Self-organized criticality The tendency of certain discrete and dissipative dynamical systems to evolve to a state where changes occur over all possible length scales.

Definition of the Subject

Self-organized criticality is a concept invoked to explain the frequent occurrence of fractal structures and multiscale phenomena in nature. In contrast with the ideas of chaos, here simple common features appear in systems with many degrees of freedom. For modeling this phenomenon, cellular automata provide an elegant class of dynamical systems which are easily simulated numerically.

Introduction

Cellular automata provide a fascinating class of dynamical systems based on very simple rules of evolution yet capable of displaying highly complex behavior. These include simplified models for many phenomena seen in nature. Among other things, they provide insight into self-organized criticality, wherein dissipative systems naturally drive themselves to a critical state with important phenomena occurring over a wide range of length and time scales.

This article begins with an overview of self-organized criticality. This is followed by a discussion of a few examples of simple cellular automaton systems, some of which may exhibit critical behavior. Finally, some of the fascinating exact mathematical properties of the Bak–Tang–Wiesenfeld sandpile model [1] are discussed.

Self-Organized Criticality

Self-organized criticality refers to the tendency of many dynamical systems to naturally drive themselves to a state displaying fluctuations over a wide range of scales [1]. The concept is invoked as a possible "explanation" of the omnipresent multi-scale structures throughout the natural world, ranging from the fractal structure of mountains, to the power law spectra of earthquake sizes [2]. Recent applications include such diverse topics as punctuated evolution [3] and traffic flow [4]. The concept has even been invoked to explain the unpredictable nature of economic systems; i. e. why you can not beat the stock market [5].

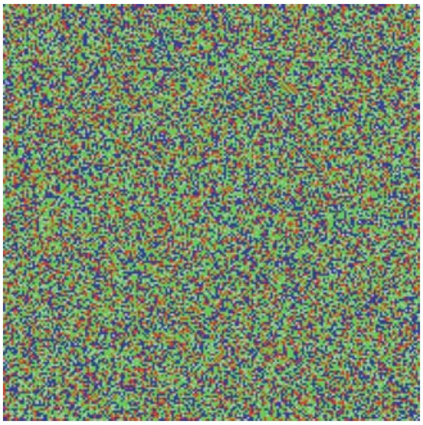
Self-organized criticality nicely compliments the concept of chaos. In the latter, dynamical systems with a few degrees of freedom, say as little as three, can display highly complex behavior, often generating beautiful fractal structures. With self-organized criticality, we start instead with systems of many degrees of freedom, and find a few general common features.

Another attractive feature of both self organized criticality and chaos is the ease with which computer models can be implemented and the elegance of the resulting graphics. Most of the figures in this chapter were produced using my publicly available set of programs "xtoys" [6]. Indeed, much of this presentation is based on my similar article in [7].

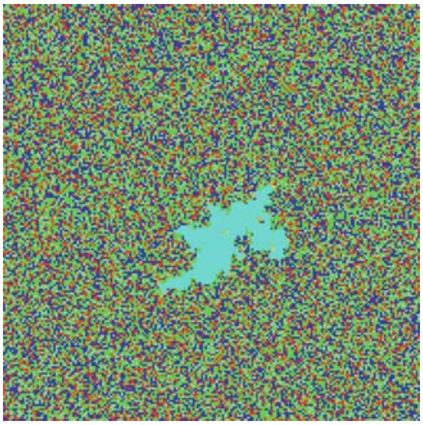
The paradigm for the phenomenon is the sandpile. On slowly adding grains of sand to an empty table, a pile will grow until its slope becomes critical and avalanches start spilling over the sides. If the slope becomes too large, a large catastrophic avalanche is likely, and the slope will reduce. If the slope is too small, then the sand will accumulate to make the pile steeper. Ultimately one should obtain avalanches of all sizes, with the prediction of the size for the next avalanche being impossible to determine without actually running the experiment.

The original Bak, Tang, Wiesenfeld paper [1] presented a particularly simple model to mimic the sandpile idea. For this, each site of a two dimensional lattice has a state represented by a positive integer z_i. This integer can be thought of as representing the amount of sand at that location, or in another sense it represents the slope of the sandpile at that point. Neither of these analogies is fully accurate, the model has aspects of each.

The dynamics follows by setting a threshold z_{T} above which any given z_i is unstable. Without loss of generality, I take this threshold to be $z_{\mathrm{T}} = 3$. Time now proceeds in discrete steps. In one such step each unstable site with $z_i \geq 4$ "tumbles" or "topples," dropping by four and adding one grain to each of its four nearest neighbors. This may produce other unstable sites, and thus an avalanche can ensue. This proceeds for further time steps until all sites are stable. Figure 1 shows a typical configuration on a 198 by 198 lattice after lots of random sand addition followed by relaxation. Figure 2 shows an avalanche proceeding on this lattice, and Fig. 3 shows the final avalanch region after the system reaches stability.

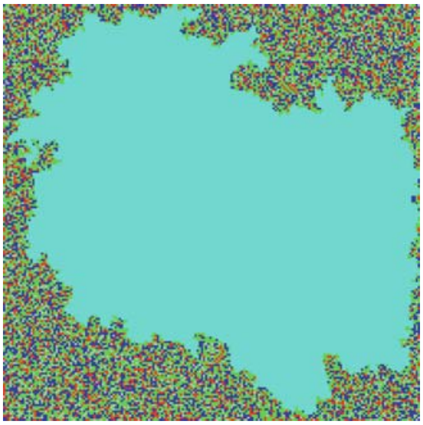


Self-organized Criticality and Cellular Automata, Figure 1
The sandpile model in the final stable state after adding lots of sand to random places. The lattice is 198 cells by 198 cells. The color code is *gray*, *red*, *blue*, and *green* for heights 0, 1, 2, and 3, respectively. Despite the lack of obvious patterns, subtle correlations are present; for example no two adjacent sites have height zero



Self-organized Criticality and Cellular Automata, Figure 2
An ongoing avalanche obtained by adding a small amount of sand to the configuration in Fig. 1. Stable sites which have tumbled during the avalanche are distinguished by being colored *light blue*. The still active sites are colored *yellowish brown*

A natural experiment consists of adding a grain of sand to a random site and measuring the number of topplings and the number of time steps for the resulting avalanche.



Self-organized Criticality and Cellular Automata, Figure 3
The final state after the avalanch in Fig. 2 has completed. The sites which tumbled during the avalanche are distinguished by being colored *light blue*. Note that the final avalanche region is simply connected. This is a general result proven later in the text

Repeating this many times to gain statistics, the distribution of avalanche sizes and lengths displays a power law behavior, with all sizes appearing. In [8] such experiments showed that the distribution of the number of tumbling events s in an avalanche empirically scales as

$$P(s) \sim s^{-1.07} \tag{1}$$

and the number of time steps τ for avalanches scales as

$$P(\tau) \sim \tau^{-1.14} \, . \tag{2}$$

This model has been extensively studied analytically. While as yet there is no exact calculation of these exponents, a lot is known. In particular, the critical ensemble is well characterized. I will return to these points later.

The extent to which laboratory experiments reproduce these phenomena is somewhat controversial. A study of avalanche dynamics [9] in rice piles showed power laws with long-grain rice, but more ambiguous results followed similar experiments with short-grain rice.

Cellular Automata

The sandpile model is a simple example of a system of cellular automata [10,11]. Each site or "cell" of our lattice follows a prescribed rule evolving in discrete time steps. At each step, the new value for a cell depends only on the current state of itself and its neighbors. These systems are fascinating in that deceptively simple rules can give rise to extremely complex behavior. Furthermore, slight changes in the rules can dramatically change their behavior.

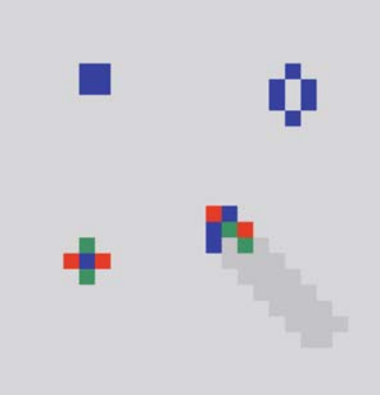
Even though the formulation of a cellular automaton may seem almost trivial, there are a huge number of possible rules. For example, suppose I consider two dimensional models where each cell can take only one of two possible states. These might be referred to as unset or set bits, or more figuratively as "dead" or "alive." Suppose furthermore that I restrict myself to rules where the evolution of a given cell to the next time step depends only on the current values of the cell and each of its eight neighbors. In this case there are $2^9 = 512$ possible arrangements for the cell and its neighbors. A general rule needs to specify the next state of the cell for each of these arrangements. This gives $2^{512} = 1.3 \times 10^{154}$ possible rules. Given that the universe is only of order 4×10^{17} seconds old, clearly only a vanishing fraction of these rules have a chance of being studied in any of our lifetimes.

A simple subset of rules called "totalistic" have the state of the updated cell only depend on the total number of living neighbors. With the eight cell neighborhood, there are nine possible values for this sum, and the new value for the cell requires specification of the new state for each of these as well as the current state of the cell. This gives $2^{18} = 262,144$ rules; still large, but not truly astronomical. If I restrict the rule to depending on the total of only the four nearest neighbors, I then have a modest $2^{10} = 1024$ cases to consider. Other than the sandpile model, most of the following will be restricted to such totalistic rules.

With a discrete set of states, cellular automata have the appealing feature of being easily implementable entirely by logical operations, the natural functions of computer circuitry. Also, the state of several cells can be stored and manipulated within a single computer word. Using such tricks, these models can often be implemented to run extremely fast, leading to hope that such models may supply simulation methods as good as or better than the conventional use of floating point fields on a discrete grid. With this motivation, considerable attention has been paid to cellular automata that may simulate fluid flow. Another advantage of this approach is the ability to work with arbitrary boundary conditions. These topics go beyond the scope of this article. A nice review can be found in [12].

Conway's Life

Perhaps the most famous cellular automaton model is Conway's "Game of Life" [13]. For this there exists a vast literature; so, I will only mention a couple of interesting features. The rule involves the eight cell neighborhood, and if a cell is initially "dead" it becomes alive if and only if it has exactly three live neighbors, or "parents". A living cell dies of loneliness if it has less than two live neighbors,



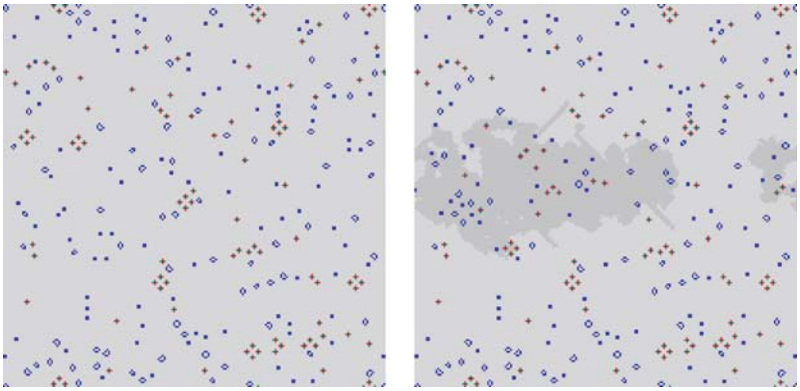
Self-organized Criticality and Cellular Automata, Figure 4
Some living configurations in life. The *top* two are stable patterns. The *lower left* shows a "blinker" or "traffic light" which oscillates with a period of two. On the *lower right* is a glider, which propagates diagonally through the lattice. *Blue* denotes a state that is and just was alive, *red* is newborn, and *green* represents just died. The track of the glider is *darkened* slightly over the remaining *gray background* to show its motion

and of overcrowding if it has more than three live neighbors. Only in the case of exactly two or three live neighbors does it survive.

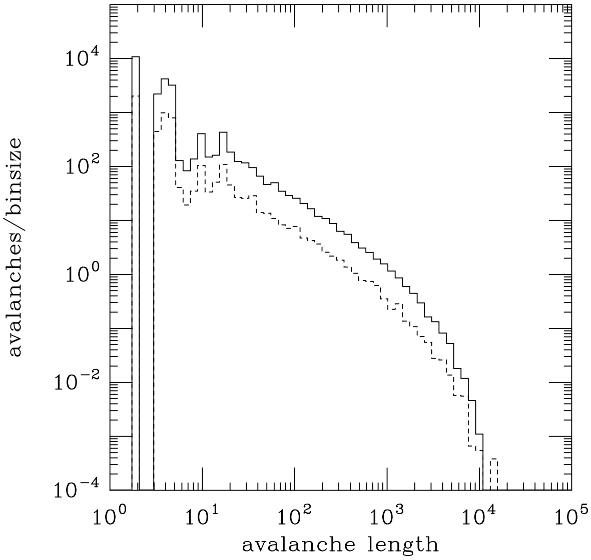
While simple to state, this model displays fascinating complexity. There are simple isolated sets of live cells that quietly survive, such as a block of four neighboring live cells forming a two by two square. Other configurations oscillate, such as three live cells in a row, which alternate between being vertically and horizontally oriented. A particularly amusing local configuration has five live cells; say starting with coordinates $\{(0,0),(0,1),(0,2),(1,2),(2,1)\}$. After four time steps this configuration returns to its original shape, but displaced by $(-1,1)$. On an otherwise empty board, this "glider" continues to propagate as a single entity. In an on-screen simulation, it appears much as a small insect crawling about. Some elementary configurations are shown in Fig. 4. A large collection of fascinating life configurations can be found in the Wikipedia [14].

Gliders allow information to be propagated over long distances, and it has been proven that with a complicated enough initial configuration, one can construct a computer out of live cells on a life board [13]. Special sub-configurations form the analog of electronic gates, which can control beams of gliders representing bits. Indeed, since life is capable of universal computation, one might imagine a life board programmed to simulate the game of life.

There is some limited evidence that the game of life also displays self-organized criticality [15,16]. One can repeatedly throw down gliders, which collide and create a background of static and oscillating clumps. While oscillators of arbitrarily long period are known to exist, those with period longer than two are extremely rare and almost never created from unorganized initialization. Once the system has settled into a loop, then another glider can be tossed on, giving a disturbance. An avalanche is defined to occur during the period until the system again goes into an oscillating state. Figure 5 shows the effect of such a disturbance. In Fig. 6 I show the distribution of such avalanches as measured on modest lattices. There is a hint of a power law superposed on additional structure from avalanches of only a few time steps, and a rounding at large times possibly due to finite size effects. The criticality of life remains controversial; [17] has looked unsuccessfully for a power



Self-organized Criticality and Cellular Automata, Figure 5
On the *left* is a configuration in life resulting from a random start and evolved until only stable and period two oscillators remain. On the *right* is the state after a small disturbance was introduced in the center and allowed to die out. Note the irregular shape of the disturbed region, which has been tinted a *darker gray*. The lattice here is 198 sites wide by 198 sites high, with periodic boundaries



Self-organized Criticality and Cellular Automata, Figure 6
The distribution of avalanches generated by adding gliders randomly to a system in the game of life consisting of stable and period two oscillators. An avalanche occupies the period until the system has relaxed again into such a periodic state. The *solid line* represents 25,000 avalanches on a 512 by 512 lattice, and the *dashed line* is for 6,000 avalanches on a 1024 by 1024 system. This figure is taken from [16]

law distribution of activity as one moves in from a source on the boundary. The relation between these two experiments is unclear.

Fredkin's Modulo-Two Rule

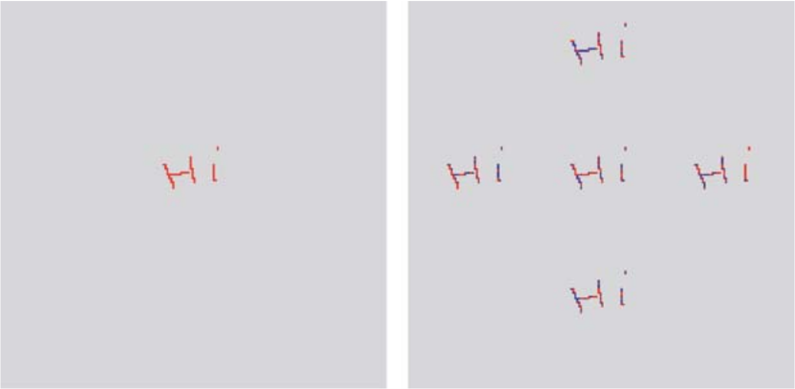
An extremely simple but highly amusing rule takes at each time step the "exclusive or" (XOR) operation between a site and its neighbors. This rule has the remarkable property of self replication [18]. Starting with any given initial pattern, after 2^n time steps copies of the original state occupy positions separated by 2^n spatial sites from the original in every direction as specified in the chosen neighborhood. In Fig. 7 I show an example of this with the four cell neighborhood.

In this rule, the pattern is generally rather complex just before returning to the replicated case, i. e. after $2^n - 1$ steps. Figure 8 shows the pattern obtained from a single set pixel after this rule has been applied for 63 time steps using the four nearest cells as the neighborhood. Note the fractal structure. In one more time step, all but five copies of the original set bit die.

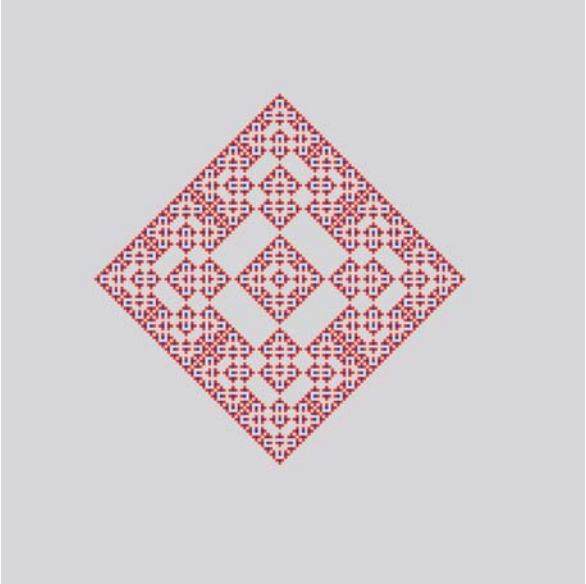
Unlike most cellular automaton rules, this gives a dynamics which in some sense is not really "complex". In most cases the simplest way to predict the evolution of a cellular automaton rule is to actually run it. Here, however, there is an easier way to predict what the final pattern will look like; it is always an XOR operation between several displaced copies of the configuration that appeared 2^n time steps in the past. Despite the lack of complexity, this rule shows rather dramatically that cellular automata are capable of "reproduction".

Reversible Rules

Reversibility is rather elusive among cellular automata. In the game of life, a single isolated cell immediately dies leaving no trace; thus it is impossible from the state at a given time to reconstruct what was there one time step back. A related difficult problem is to construct "garden of Eden" configurations which are impossible to arrive at from any previous state [19].



Self-organized Criticality and Cellular Automata, Figure 7
Starting from the initial configuration *on the left*, the modulo two rule is evolved for 64 steps using the four nearest neighbors. At a certain stage, five copies of the original image appear. The *blue pixels* indicate which sites were also alive one step before



Self-organized Criticality and Cellular Automata, Figure 8
The state after applying 63 steps of the modulo two rule using the four nearest neighbors to an initial state of a single set bit. After the next time step this fractal structure decays into only five remaining live cells

Fredkin pointed out an interesting class of reversible rules based on an analogy with molecular dynamics [11]. In the later one specifies both the position and the velocities of a set of particles and evolves the system under Newton's equations with some given inter-particle force law. Reversal can then be accomplished by merely changing the signs of all the velocities.

In a cellular automaton an analog of velocity requires the value of the cells at two successive time steps. Based on this, Fredkin presented a very simple scheme using the previous state to generate a wide class of reversible rules. He considered taking an arbitrary automaton rule at a given time, and then added an exclusive or (XOR) operation of the result with the state one step back in time. These combined operations could then be reversed by merely interchanging two successive time steps, the analogy of reversing the velocities.

To see this more mathematically, suppose the state at time i is s_i, and the underlying rule begins by taking some arbitrary function $f(s_i)$. Then the full rule takes for the next time step $s_{i+1} = f(s_i)$ XOR s_{i-1}. Here the exclusive or operation is taken site by site over the entire lattice. Elementary properties of the XOR operation then give $s_{i-1} = f(s_i)$ XOR s_{i+1}, which is the identical rule for the time reversed dynamics.

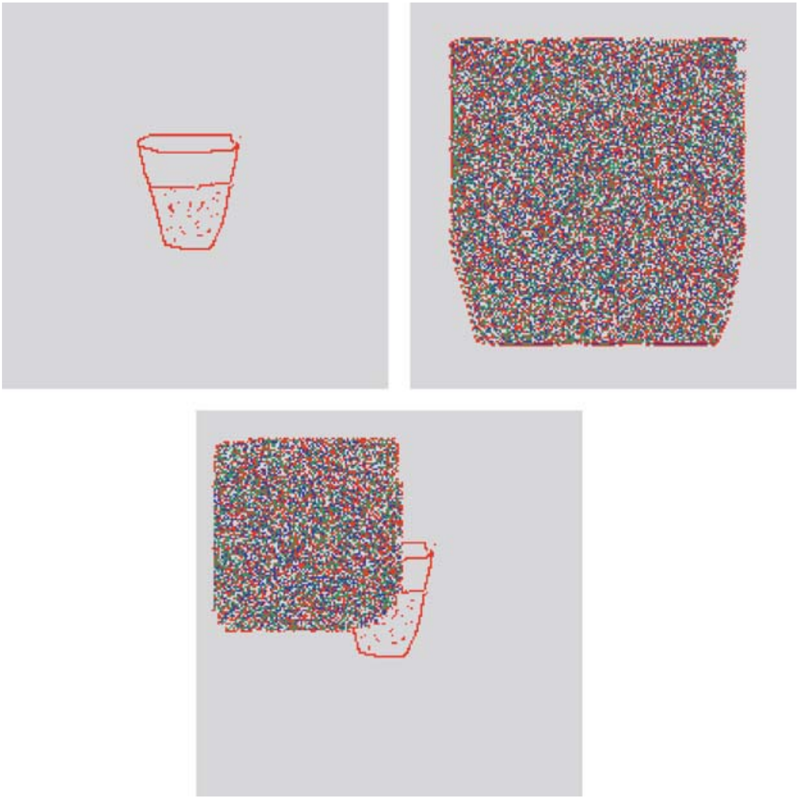
These rules provide a wonderful way to play with the concepts of entropy and reversibility. Indeed, an idealized universe of cellular automata enables experiments which would be impossible to carry out in the real world. In Fig. 9 I show the evolution of a simple image under such a rule. The experiment is a crude simulation of a beer glass shattering after being dropped on the floor. After a few steps it appears quite randomized. Reversal of the momenta of all relevant atoms in the beer glass would allow its reconstruction. In the model this is easily accomplished by swapping two time steps. After reversal, continuing with the same rule reconstructs the original image. At all stages the "information" contained in the system must be constant, even though the image may appear of drastically different complexity.

The reconstruction process is highly sensitive to the reversal being precise. The analog here is to the sensitivity to initial conditions in dynamical systems. In the bottom of Fig. 9c I try to reproduce the beer cup from its shards as in the above experiment, except that now at the time of reversal I modify the state of exactly one pixel. The reversal process recovers the original image only in regions outside the "light cone" for the modified pixel. As the disturbance can only propagate to neighbors in one time step, pixels outside n steps can not know of the change before an equal number of time steps. This use of an XOR operation to generate reversible complex mappings is an integral part of the Data Encryption Standard; see, for example, [20].

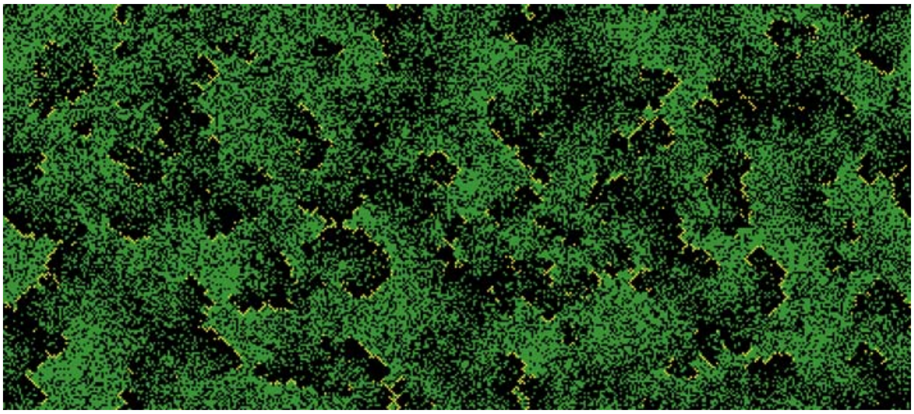
Forest Fires

An amusing model of forest fires has three possible states per cell, empty, a tree, or a fire. For the updating step, any empty site can have a tree born with a small probability. At the same time, any existing fire spreads to neighboring trees leaving its own cell empty. The rule here differs from those discussed previously in having a stochastic nature. As the system is made larger, the growth rate for the trees should decrease to just enough to keep the fires going.

If too many trees grow, one obtains a large fire reducing their density, while if there are too few trees, fires die out. On a finite system, one should light a fire somewhere to get the system started. On the other hand, as the system becomes larger, the growth rate for the trees can be reduced without the fire expiring. In a steady state the system has fire fronts continually passing through the system, as illustrated in Fig. 10. Perhaps there is a moral here that one should be careful about extinguishing all fires in the real world, for this may enhance the possibility for a catastrophic uncontrollable fire. It is not entirely clear whether this model is actually critical. What seems to happen on



Self-organized Criticality and Cellular Automata, Figure 9
The encryption of a glass of beer. The original rule uses the eight cell neighborhood with births on 1, 3, 5, and 7 neighbors and survivors on exactly 1 neighbors. The rule is modified at each step by XOR'ing the result with the history one time step back. Swapping two adjacent time steps will bring the glass back exactly. The first figure is the starting configuration, the second after 50 steps of evolution. At this point one bit in the *upper left hand quadrant* is flipped, and the dynamics is reversed. The glass is restored in all places beyond 50 steps from the flipped bit. Note the effect of a "speed of light" in the problem



Self-organized Criticality and Cellular Automata, Figure 10
A snapshot of the forest fire model on a 450 by 200 periodic lattice. Trees are continuously burning at a slow rate, while fires burn them down and spread to nearest neighbor trees. Here the four cell neighborhood is used

large systems is that stable spiral structures form and set up a steady rotation. For a review of this and several related models, see [21].

The Sandpile Revisited

Very little is known analytically about general cellular automata. However, in a series of papers, Deepak Dhar and co-workers have shown that the sandpile model has some rather remarkable mathematical properties [22,23,24,25]. In particular, the critical ensemble of the system has been well characterized in terms of an Abelian group. In the following I will generally follow the discussions given in [2,26].

Dhar introduced the useful toppling matrix $\Delta_{i,j}$ with integer elements representing the change in the height, z at site i resulting from a toppling at site j [22]. More precisely, under a toppling at site j, the height at any site i becomes $z_i - \Delta_{i,j}$. For the simple two dimensional sand model the toppling matrix is thus

$$\begin{array}{ll} \Delta_{i,j} = 4 & i = j \\ \Delta_{i,j} = -1 & i, j \text{ nearest neighbors} \\ \Delta_{i,j} = 0 & \text{otherwise.} \end{array} \tag{3}$$

For this discussion there is little special to the specific lattice geometry; indeed, the following results easily generalize to other lattices and dimensions. The analysis requires only that under a toppling of a single site i, that site has its slope decreased ($\Delta_{i,i} > 0$), the slope at any other site is either increased or unchanged ($\Delta_{i,j} \le 0$, $j \ne i$), the total amount of sand in the system does not increase ($\sum_j \Delta_{i,j} \ge 0$), and, finally, that each site be connected through toppling events to some location where sand can be lost, such as at a boundary.

For the specific case in Eq. (3), the sum of slopes over all sites is conserved whenever a site away from the lattice edge undergoes a toppling. Only at the lattice boundaries can sand be lost. Thus the details of this model depend crucially on the boundaries, which we take to be open. A toppling at an edge loses one grain of sand and at a corner loses two.

The actual value of the maximum stable height z_T is unimportant to the dynamics. This can be changed by simply adding constants to all the z_i. Thus, as in Sect. "Self-Organized Criticality", I consider $z_T = 3$. With this convention, if all z_i are initially non-negative they will remain so, and I thus restrict myself to states C belonging to that set. The states where all z_i are non-negative and less than 4 are called stable; a state that has any z_i larger than or equal to 4 is called unstable. One conceptually useful configuration is the minimally stable state C^* which has all the heights at the critical value z_T. By construction, any addition of sand to C^* will give an unstable state leading to a large avalanche.

I now formally define various operators acting on the states C. First, the "sand addition" operator α_i acting on any C yields the state $\alpha_i C$ where $z_i = z_i + 1$ and all other z are unchanged. Next, the toppling operator t_i transforms C into the state with heights z'_j where $z'_j = z_j - \Delta_{i,j}$. The operator U which updates the lattice one time step is now simply the product of t_i over all sites where the slope is unstable,

$$UC = \prod_i t_i^{p_i} C \tag{4}$$

where $p_i = 1$ if $z_i \ge 4$; 0 otherwise. Using U repeatedly gives the relaxation operator R. Applied to any state C this corresponds to repeating U until no more z_i change. Neither U nor R have any effect on stable states. Finally, I define the avalanche operators a_i describing the action of adding a grain of sand followed by relaxation

$$a_i C = R \alpha_i C\,. \tag{5}$$

At this point it is not entirely clear that the operator R exists; in particular, it might be that the updating procedure enters a non-trivial cycle consisting of a never ending avalanche. I now prove that this is impossible. First note that a toppling in the interior of the lattice does not change the total amount of sand. A toppling on the boundary, however, decreases this sum due to sand falling off the edge. Thus, during an avalanche the total sand in the system is a non-increasing quantity. No closed cycle can have toppling at the boundary since this will decrease the sum. Next, the sand on the boundary will monotonically increase if there is any toppling one site further in. This also can not happen in a cycle; thus, there can be no topplings one site away from the edges. By induction there can be no toppling arbitrary distances in from the boundary; thus, there can be no cycle, and the relaxation operator exists. Note that for a general geometry this argument requires that every site be eventually connected to an edge where sand can be lost.

With a system lacking edges, such as under periodic boundaries, no sand would be lost and thus cycles are expected and easily observed. These models might be called "Escher models" after the artist constructing drawings of water flowing perpetually downhill and yet circulating in the system. While little is known about the dynamics of this variation on the sandpile model, some studies have been done under the nomenclature of "chip-firing games" [27]. It has been argued [28] that this lossless sandpile model on an appropriate lattice is capable of universal computation.

I now introduce the concept of recursive states. This set, denoted $\mathcal{R}$, includes those stable states which can be reached from any stable state by some addition of sand followed by relaxation. This set is not empty because it contains at least the minimally stable state C^*. Indeed, that state can be obtained from any other by carefully adding just enough sand to each site to make each z_i equal to three. Thus, one might alternatively define $\mathcal{R}$ as the set of states which can be obtained from C^* by acting with some product of the operators a_i.

It is easily shown that there exist non-recursive, transient states; for instance, no recursive state can have two adjacent heights both being zero. If you try to tumble one site to zero height, then it drops a grain of sand on its neighbors. If you then tumble a neighbor to zero, it dumps a grain back on the original site. One can also show that the self-organized critical ensemble, reached under random addition of sand to the system, has equal probability for each state in the recursive set. This is a consequence of the Abelian nature of this system, as discussed below.

The crucial results of [22,23,24,25] are that the operators a_i acting on stable states commute, and they generate an Abelian group when restricted to recursive states. I begin by showing that the operators commute, that is $a_i a_j C = a_j a_i C$ for all C. First I express the a's in terms of toppling and adding operators

$$a_i a_j C = \left(\prod_{k=1}^{n_1} t_{l_k}\right) \alpha_i \left(\prod_{k=n_1+1}^{n} t_{l_k}\right) \alpha_j C \qquad (6)$$

where the specific number of topplings n_1 and n depend on i, j, and C. Acting on general states, the operators t and α all commute because they merely linearly add or subtract heights. Therefore I can shift α_i to the right in this expression:

$$a_i a_j C = \left(\prod_{k=1}^{n} t_{l_k}\right) \alpha_i \alpha_j C\,. \qquad (7)$$

Now I rearrange the product of topplings. In the nontrivial case that the α-operators render either i or j (or both) unstable, the product must contain toppling operators corresponding to those unstable sites. I shift those operators to the right. Those operators constitute by definition the update operator, U, so I can write

$$a_i a_j C = \left(\prod t_{l_k}\right) U \alpha_i \alpha_j C \qquad (8)$$

where the factors within the bracket are the remaining t's. Now, the update operator may leave some sites still unstable, and then the product must include further toppling operators; working on those sites, I can pull out another factor of the update operator. This procedure can be repeated until I have used all the toppling factors and the state is stable. Thus, I can identify the operator within the brackets in Eq. (8) as the relaxation operator R. But $\alpha_i \alpha_j C$ is the same state as $\alpha_j \alpha_i C$, so $a_i a_j C = a_j a_i C$.

A trivial consequence of this argument is that the total number of tumbling events occurring in the operations $a_i a_j C$ and $a_j a_i C$ are the same. Of course, if a particular site k tumbles it can be caused by either addition; the orders of the tumbling events may or may not be altered.

An intuitive argument that sand addition may be commutative uses an analogy with combining many digit numbers under long addition. The tumbling operation is much like carrying, except rather than transferring to the next digit, the overflow spreads to several neighbors. As addition is known to be Abelian, despite the confusing elementary-school rules, I might expect the sandpile addition rule also to be.

I now prove that the avalanche operators have unique inverses when restricted to recursive states; that is, there exists a unique operator a_i^{-1} such that $a_i(a_i^{-1}C) = C$ for all C in $\mathcal{R}$. This implies that the operators a_i acting on the recursive set generate an Abelian group. For any recursive state C I first find another recursive state such that a_i acting on it gives C, and I then show that this construction is unique.

I begin by adding a grain of sand at site i to the state C and then relax the system. This generates a new recursive state $a_i C$. Now since the state C is by assumption recursive, there is some way to add sand to regenerate C from any given state. In particular, there is some product P of addition operators a_j such that

$$C = P a_i C\,. \qquad (9)$$

But the a's commute, so I have

$$C = a_i P C \qquad (10)$$

and thus PC is a recursive state on which a_i gives C.

I must now show that this state is unique. Consider repeating the above process to find a series of states C_n satisfying

$$(a_i)^n C_n = C\,. \qquad (11)$$

Because on a finite system the total number of stable states is finite, the sequence of states C_n must eventually enter a loop. I can run backwards around this loop by adding back the sand repeatedly to the given site. As the original state C appears in resupplying the sand, C itself must itself belong to the loop. Calling the length of the loop m, I have $(a_i)^m C = C$. I now uniquely define $a_i^{-1} C = a_i^{m-1} C$.

I now have sufficient machinery to count the number of recursive states. As all such can be obtained by adding sand to C^*, I can write any state $C \in \mathcal{R}$ in the form

$$C = \left(\prod_i a_i^{n_i} \right) C^* \,. \tag{12}$$

Here the integers n_i represent the amount of sand to be added at the respective sites. However, in general there are several different ways to reach any given state. In particular, adding four grains of sand to any one site must force a toppling and is equivalent to adding a single grain to each of its neighbors. This can be expressed as the operator statement

$$a_i^4 = \prod_{j \in nn} a_j \tag{13}$$

where the product is over the nearest neighbors to site i. I can rewrite this equation by multiplying by the product of inverse avalanche operators on the nearest neighbors on both sides, thus obtaining for any site i

$$\prod_j a_j^{\Delta_{ij}} = E \tag{14}$$

where E is the identity operator. This allows me to shift the powers appearing in Eq. (12). Define N to be the number of sites in the system. If I label states by the vector $\mathbf{n} = (n_1, n_2, n_3, \dots, n_N)$ I see that two states are equivalent if the difference of these vectors is of the form $\sum_j \beta_j \Delta_{ij}$ where the coefficients β_j are integers. These are the only constraints; if two states can not be related by toppling they are independent. Thus any vector $\mathbf{n}$ can be translated repeatedly until it lies in an N-dimensional hyper-parallelepiped whose base edges are the vectors Δ_{ji}, $j = 1, \dots, N$. The vertices of this object have integer coordinates and its volume is the number of integer coordinate points inside it. This volume is just the absolute value of the determinant of Δ. Thus the number of recursive states equals the absolute value of the determinant of the toppling matrix Δ.

For large lattices this determinant can be found easily by Fourier transform. In particular, whereas there are $4N$ stable states, there are only

$$\exp\left(N \int_{(-\pi,-\pi)}^{(\pi,\pi)} \frac{\mathrm{d}^2 q}{(2\pi)^2} \ln(4 - 2q_x - 2q_y) \right) \simeq (3.2102\dots)^N \tag{15}$$

recursive states. Thus starting from an arbitrary state and adding sand, the system "self-organizes" into an exponentially small subset of states forming the attractor of the dynamics.

An Isomorphism

Following [26], I now look into the consequences of stacking sand piles on top of one another. Given stable configurations C and C' with configurations z_i and z'_i, I define the state $C \oplus C'$ to be that obtained by relaxing the configuration with heights $z_i + z'_i$. Clearly, if either C or C' is a recursive state, so is $C \oplus C'$.

Under the operation $\oplus$ the recursive states form an Abelian group isomorphic to the algebra generated by the a_i. First, the addition of a state C with heights z_i is equivalent to operating with a product of a_i raised to z_i, that is

$$B \oplus C = \left(\prod a_i^{z_i} \right) B \tag{16}$$

for any recursive state B. The operation $\oplus$ is associative and Abelian because the operators a_i are.

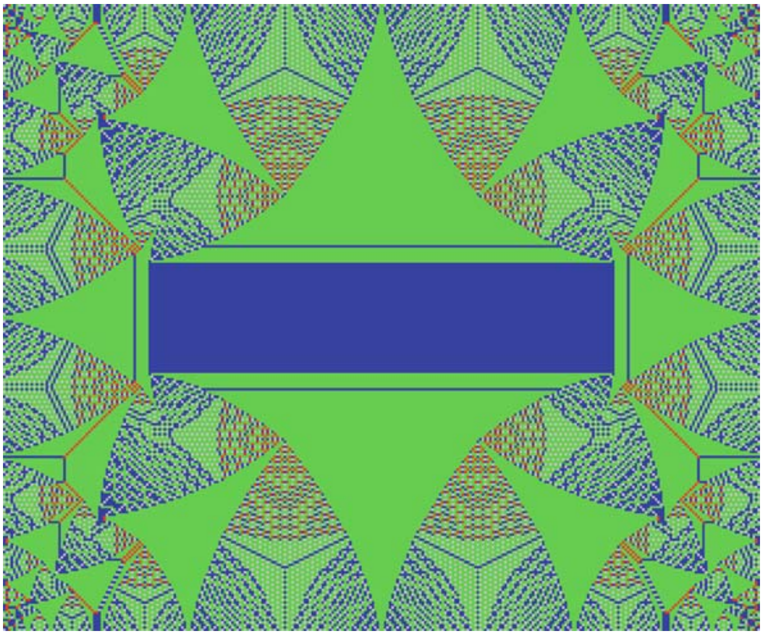
Since any element of a discrete group raised to the order of the group gives the identity, it follows that $a_i^{|\Delta|} = E$. This implies the simple formula $a_i^{-1} = a_i^{|\Delta|-1}$. The analog of this for the states is the existence of an inverse state, $-C$

$$-C = (|\Delta| - 1) \otimes C \,. \tag{17}$$

Here, $n \otimes C$ means adding n copies of C and relaxing. The state $-C$ has the property that for any state $B \oplus C \oplus (-C) = B$.

The state $I = C \oplus (-C)$ represents the identity and has the property $I \oplus B = B$ for every recursive state B. The state which is isomorphic to the operator a_i is simply $a_i I$. The identity state provides a simple way to check if a state, obtained for instance by a computer simulation, has reached the attractor, i. e. if a given state is a recursive state: A stable state is in $\mathcal{R}$ if and only if $C \oplus I = C$. The proof is simple. By construction, a recursive state has this property. On the other hand, since I is recursive, so is $C \oplus I$.

The identity state can be constructed by taking any recursive state, say C^* and repeatedly adding it to itself to use $|\Delta| \otimes C = I$. However, on any but the smallest lattices, $|\Delta|$ is a very large integer. A more economical scheme is to start with an empty table but use sandy boundary conditions which continually pour sand onto the table. Once it reaches a steady state, switch to open boundary condi-

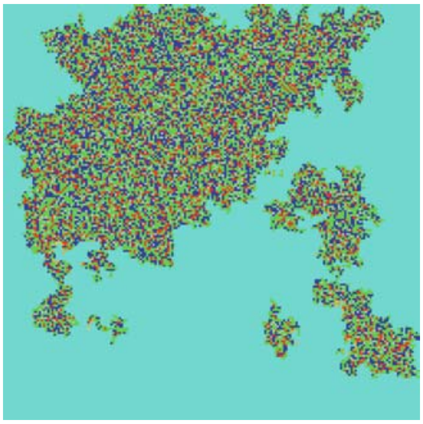


Self-organized Criticality and Cellular Automata, Figure 11
The identity state for the sandpile model on a 302 by 250 lattice. The color code is *gray*, *red*, *blue*, and *green* for heights 0, 1, 2, and 3, respectively

tions and let the sand run back off. This then relaxes to the desired identity. Figure 11 shows the identity state on a 302 by 250 lattice. Note the fractal structure, with features on many length scales.

Majumdar and Dhar [25] have constructed a simple "burning" algorithm to determine if a state belongs to the recursive set. For a given configuration, first add one particle to each of the edge sites and two particles to the corners. This again corresponds to imagining a large source of sand just outside the boundaries, which then tumbles one step onto the system. Then return to open boundaries and update according to the usual rules. If and only if the original state is recursive, this will generate an avalanche under which each site of the system tumbles exactly once. Also, the final state after the avalanche will be identical to the original. However, if the state is not recursive, some untumbled sites will remain. Figure 12 shows such a process underway on the configuration of Fig. 1. Here sites which have already burned are shown in cyan, while the remaining sites in the center have not yet tumbled. The small number of sites shown in orange are the still active sites, which eventually burn the entire remaining lattice.

The burning algorithm provides a simple way to prove that the avalanche regions are simply connected once one is in the critical state. In a burning process, any sub-lattice of the original will have all of its sites tumbled onto from outside. This is the condition for starting a burning on the sub-lattice. Thus, if a configuration is in the critical ensemble for the whole lattice, then any extracted piece of



Self-organized Criticality and Cellular Automata, Figure 12
The burning algorithm being applied to the state in Fig. 1. Burnt sites are *cyan*, burning sites are *orange*, and the remaining sites are colored as previously. This avalanche eventually tumbles every site exactly once

this configuration on a subset of the original lattice is also in the critical ensemble of the extracted part. Now suppose that one constructs an avalanche with any initial addition to a state from the critical ensemble. In any subregion enclosed by this avalanche, sand will fall from the tumbling sites on its outside. Since the sub-lattice is itself in its own critical ensemble, this must induce an avalanche which, by the burning algorithm, will tumble all enclosed sites. Thus any avalanche on a state from the critical ensemble cannot leave untumbled any sites in a region isolated from the boundary, i. e. an untumbled island. This result that avalanches must be simply connected does not follow for states outside the recursive set, as can be easily demonstrated by considering a sandpile with a hole of empty sites in the middle.

The burning algorithm has several amusing consequences. One is that any configuration with only height 2 or 3 present is in the critical ensemble as long as the lattice has corners. For example, with all height 2, the burning will start at the four corners of a rectangular lattice and steadily work its way to the center of the system. Another consequence is that in addition to the tumbling region from an avalanche being simply connected, so will the smaller region where the number of tumblings exceed any fixed number; i. e. the region of sites that tumble twice or more is also simply connected.

Future Directions

Simple models as implemented by cellular automata provide a rich area for the study of complex phenomena. Some systems can self organize with physics at many scales, while others provide fascinating demonstrations of thermodynamic laws. I have only touched on a few issues here, leaving out many related topics such as lattice gasses, driven interfaces in random media, growth processes, and evolution. As the ease of programming and the speed of modern computers continue to rush forward, so will the fascination with such models.

Acknowledgments

I am thankful for discussions with many people, but most particularly P. Bak, D. Dar, E. Fredkin, N. Margolis, M. Paczuski, T. Toffoli, F. Van Scoy, and G. Vichniac. This manuscript has been authored under contract number DE-AC02-76CH00016 with the US Department of Energy. Accordingly, the US Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for US Government purposes.

Bibliography

Primary Literature

1. Bak P, Tang C, Wiesenfeld K (1987) Phys Rev Lett 59:381; (1988) Phys Rev A 38:3645
2. Bak P, Creutz M (1994) Fractals and self-organized criticality. In: Bunde A, Havlin S (eds) Fractals in Science. Springer, Berlin, pp 26–47
3. Paczuski M, Maslov S, Bak P (1996) Phys Rev E 53:414
4. Nagel K, Paczuski M (1995) Phys Rev E 51:2909
5. Levy M, Solomon S, Ram G (1996) Int J Mod Phys C 7:65
6. The latest version of the xtoys package is available at http://thy.phy.bnl.gov/www/xtoys/xtoys.html
7. Creutz M (1997) Cellular automata and self organized criticality. In: Bhanot G, Chen S, Seiden P (eds) Some new directions in science on computers. World Scientific, Singapore, pp 147–169
8. Christensen K (1992) Ph D Thesis, University of Aarhus
9. Frette V et al (1996) Nature 379:49
10. Wolfram S (1986) Theory and Applications of Cellular Automata. World Scientific, Singapore
11. Toffoli T, Margolus N (1987) Cellular Automata Machines. MIT Press, Cambridge
12. Bogosian B (1993) Nucl Phys B, Proc Suppl 30:204
13. Berlekamp E, Conway J, Guy R (1982) Winning Ways for your Mathematical Plays, vol 2. Academic Press, New York
14. Wikipedia (2007) Conway's Game of Life. http://en.wikipedia.org/wiki/Conway's_life. Accessed 6 Apr 2007
15. Bak P, Chen K, Creutz M (1989) Nature 342:780
16. Creutz M (1992) Nuclear Phys B, Proc Suppl 26:252
17. Bennett C, Bourzutschy M (1991) Nature 350:468
18. Gardner M (1983) Wheels, Life, and Other Mathematical Amusements. W.H. Freeman, New York
19. Wikipedia (2007) Garden of Eden pattern. http://en.wikipedia.org/wiki/Garden_of_Eden_pattern. Accessed 6 Apr 2007
20. Press W, Teukolsky S, Vetterling W, Flannery B (1988) Numerical Recipes in C. Cambridge University Press, Cambridge
21. Clar S, Drossel B, Schwabl F (1996) Phys J Cond Mat 8:6803
22. Dhar D (1990) Phys Rev Lett 64:1613
23. Dhar D, Ramaswamy R (1989) Phys Rev Lett 63:1659
24. Dhar D, Majumdar SN (1990) Phys J A 23:4333
25. Majumdar SN, Dhar D (1992) Physica A 185:129
26. Creutz M (1991) Comp Phys 5:198
27. Anderson R et al (1989) Amer Math Monthly 96:981; Björner A, Lovász L, Shor P (1991) Europ J Combinatorics 12:283; Eriksson K (1996) SIAM J Discret Math 9:118
28. Goles E, Margenstern M (1996) Int J Mod Phys C 7:113

Books and Reviews

Bak P (1996) How Nature Works: The Science of Self-Organised Criticality. Springer, Berlin

Gore A (1992) Earth in the Balance: Ecology and the Human Spirit. Plume, Boston

Jensen HJ (1998) Self-Organized Criticality: Emergent Complex Behavior in Physical and Biological Systems. Cambridge, Cambridge

Toffoli T, Margolus N (1987) Cellular Automata Machines: A New Environment for Modeling. MIT Press, Cambridge

Wolfram S (1994) Cellular Automata and Complexity: Collected Papers. Westview Press, Boulder

Self-organizing Systems

WOLFGANG BANZHAF
Department of Computer Science, Memorial University of Newfoundland, St. John's, Canada

Article Outline

Glossary

Attractor A special set of system states approached by a dynamical system after some time has passed when starting from a variety of initial states.

Autopoiesis The process by which systems maintain their identity and organization and regenerate their components in the course of their operation.

Competition and cooperation Types of interaction between two or more elements of a system. Competition refers to each element striving to maximize its use of a finite and/or non-renewable resource. Cooperation refers to the elements engaging in a mutually beneficial exchange.

Complexity Measure of number of elements and way of their interaction (structural c.); measure of variety of behavioral repertoire of a system (functional c.).

Constructive system A system whose later components are generated during the interaction of its earlier components.

Dynamics The quantitative development of a system's state variables over time.

Emergence The appearance of qualitatively new phenomena on higher levels of a hierarchical system.

Evolution A process of structural or qualitative change in some direction.

Instability Inability of a system to keep its state or structure.

Mode Macroscopic behavior of a system caused by the interaction of its microscopic parts via long-range correlations.

Non-equilibrium System state with inflow of matter, energy and/or information causing it to stay away from its most probable state under the hypothetical condition of isolation.

Phase transition A point at which the appearance or behavior, or qualitative nature of the steady state of a system changes suddenly.

Resilience Measure of a system's ability to remain within a domain of stability in response to fluctuations of the system by a perturbation, and the ability of the system to return to that stable domain having once left.

Self-organized criticality The ability of a system to evolve in such a way as to approach a critical point and then maintain itself at that point.

Definition of the Subject

Self-organization is a core concept of Systems Science. It refers to the ability of a class of systems (self-organizing systems (SOS)) to change their internal structure and/or their function in response to external circumstances. Elements of self-organizing systems are able to manipulate or organize other elements of the same system in a way that stabilizes either structure or function of the whole against external fluctuations. The process of self-organization is often achieved by growing the internal space-time complexity of a system and results in layered or hierarchical structures or behaviors. This process is understood not to be instructed from outside the system and is therefore called *self-organized*.

Modern ideas about self-organization start with the foundation of cybernetics in the 1940s. W. Ross Ashby, H. von Foerster and N. Wiener, among others, have contributed to an early understanding. Later, the concept was adopted in physics and nowadays pervades most of natural sciences. Many systems have been identified as possessing aspects of self-organization, though a clear definition is still lacking. As a result of this inaccuracy, the theory of self-organization is still in its infancy. While the concept has found applications in the social sciences and engineering as well, SOSs are an area of active research, with fundamental questions still being explored.

Introduction

Over the last decades a variety of features have been identified as typical for self-organizing systems. Not all of these features are present in all systems able to self-organize. Self-organizing systems are *dynamic*, often *non-determin-*

istic, *open*, exist *far from equilibrium* and sometimes employ *autocatalytic amplification of fluctuations*. Often, they are characterized by *multiple time-scales* of their internal and/or external interactions, they possess a *hierarchy of structural and/or functional levels* and they are able to *react* to external input *in a variety of ways*. Many self-organizing systems are *non-teleological*, i. e. they do not have a specific purpose except their own existence. As a consequence, *self-maintenance* is an important function of many self-organizing systems. Most of these systems are *complex* and use *redundancy* to achieve *resilience* against external pertubation tendencies.

Key aspects of self-organizing systems are:

- Growth of Complexity
- Emergence of new phenomena
- Positive and negative feedback loops of internal regulation.

The process of self-organization has been invoked to explain numerous phenomena in the natural sciences. From non-living systems like galaxies and stars down to nanoparticle aggregates, self-organizing systems have been observed. In the living world cells, organisms and ecosystems provide examples of systems classified as self-organizing. The concept has found applications in man-made systems like communication networks, societies, economies, and has been identified to be at work in the world of ideas in the development of world views, scientific beliefs and norm systems.

History of the Concept of Self-Organization

Early History

The concept of self-organization can be traced back to at least two sources: Western philosophy influenced heavily by Greek thinking; and eastern philosophy, centered around the process thinking of Bhuddism. The ideas derived from both sources resound with the modern way of thinking about self-organization although the word itself had never been used.

On wondering about the origin of the world, Greek atomists from Democritos of Abdera to Epicuros of Samos argued that world order arose from chance collisions of particles. First, the cosmos (from Greek *kosmos* = *the ordered*) did not exist but chaos instead (from Greek *chaos* = *the disordered*). In modern times chaos theory has taken up this topic again, with deep connections to ideas about self-organization and the origin of order in the universe.

In the Christian tradition, St Thomas Aquinas contributed through his interest in logical proofs for the existence of God. One of these proofs considered God to be the ultimate organizer or designer. The argument was that everything had to be organized and this called for an organizer. In turn, the organizer had to be organized and so on back to the original organizer: this was God. Since God is present without cause (otherwise he would have to be organized by another entity), he must have somehow organized himself.

The Bhuddist way of thinking, on the other hand, was fundamentally process-oriented. Things are considered not to be in static existence, but rather are thought to be generated and maintained by proper processes. The emphasis on processes is reminescent of self-organizing systems whose structure is determined by proper processes of internal and external interactions.

The First Use of the Term

Work on General Systems Theory (von Bertalanffy) [1] and Cybernetics (Wiener) [2] paved the way for the idea of self-organization.

The concept of a self-organizing system was introduced by Ashby in 1947 [3]. In the 1950s a self-organizing system was considered to be a system which changes its basic structure as a function of its experience and environment. The term was used by Farley and Clark in 1954 to describe learning and adaptation mechanisms [4]. Ashby [5], in 1960, redefined a self-organizing system to include the environment with the system proper. Von Foerster argued [6], also in 1960, that a self-organizing dynamical system possesses some stable structures (eigenvalues, attractor states) which he later termed eigenbehavior.

Further Developments

This notion was further developed by Haken [7] in 1977 who termed the global cooperation of elements of a dynamical system – resulting in it assuming an attractor state – *self-organization*. Both Haken and Kauffman (1993) [8] argued for a deep connection between self-organization and *selection*. Haken found that modes of collective behavior are competing against each other and considered this process to be Darwinian selection in the non-living world. Kauffman, on the other hand, emphasized the role of constraints on the direction of evolution (mostly of the living), caused by self-organization.

Already in the 1970s, however, ideas branched out into different directions. One branch of the development of the idea deepened the relation to studies of learning and adaptation (Conrad, Kohonen, [9,10]), another branch studied processes of self-organization in systems far from equilibrium (Prigogine, Haken) [11,12]. Chaos

Self-organizing Systems, Table 1
Examples of Self-Organizing Systems

System	Flow	Self -organizing entities	Emergence
Atmosphere	Solar energy	Gas molecules	Patterns of atmospheric circulation
Climate	Energy	Weather conditions (humidity, precipitation, temperature, ...)	Distribution patterns
Liquid between plates	Heat	Particle circulation	Movement patterns
Laser	Excitation energy	Phase of light waves	Phase-locked mode
Reaction vessel	Chemicals for BZ reaction	Chemical reactions	Patterns of reaction fronts
Neural networks	Information	Synapses	Connectivity patterns
Living cells	Nutrients	Metabolic reactions	Metabolic pathways/network patterns
Food webs	Organisms of different species	Species rank relation	Networks of species
Highway traffic	Vehicles	Distance of vehicles	Density waves of traffic
City	Goods, information	Human housing density	Settling patterns
Internet	Computer nodes	Connections between nodes	Network connection pattern
Web	Information posted in websites	Links between websites	Patterns of web communities

theory (Thom, Mandelbrot) [13,14] was the line of inquiry into nonlinear systems in mathematics, whereas autopoiesis and self-maintenance where at center stage in biology (Eigen, Rosen) [15,16] neurophysiology (von der Mahlsburg, Linsker [17,18]) and cognitive science (von Foerster, Maturana and Varela) [19,20].

In recent years, self-organizing systems have assumed center stage in the natural sciences [21,22], and the social sciences [23,24,25]. Engineering is beginning to see the usability of the concept [26] in connection with the approach of nano-scale applications and the growing complexity of human artefacts.

Examples of Natural Self-Organizing Systems

Classical examples of natural self-organizing systems are the formation of Benard convection cells in non-equilibrium thermodynamics, the generation of laser light in non-linear optics and the Belousov–Zhabotinsky reaction in chemistry. These are examples from the non-living world, and the complexity of resulting macroscopic space-time patterns is restricted.

Nearly unrestricted complexity through self-organization can be achieved in the living world. For instance, the interaction of species in foodwebs could be looked at from this point of view [22]. Here, we shall briefly look at the self-organization of the Earth's biosphere known as the Gaia hypothesis [27]. This hypothesis states that Earth's living and non-living components self-organize into a single entity called *Gaia*. Gaia can be understood as the whole of the biosphere, that is able to self-stabilize. The model states, in other words, that the biomass of Earth self-regulates to make conditions on the planet habitable for life. In this way, a sort of homeostasis would be sought by the self-organizing geophysical/physiological system of Earth.

In recent years, the Gaia hypothesis has found its place in Earth Systems Science as the realization that there is just one global ecosystem, containing the entirety of resources and all living organisms, all interacting with each other in multiple regulatory cycles. These ideas have been connected to the Darwinian theory of evolution via natural selection [28,29], providing a mechanism by which such a stable state can be assumed to have emerged.

Other examples of natural self-organizing systems can be found in Table 1.

Examples of Artificial Self-Organizing Systems

There are numerous examples of man-made systems or systems which involve man that exhibit self-organization phenomena. Among them are traffic patterns, self-organizing neural networks, celular phone networks or the development of web communities.

The example we shall briefly discuss is that of traffic flow patterns. Macroscopic patterns of traffic jams on highways have been observed and experimentally examined [30]. Their appearance is closely related to traffic density, the model of behavior for drivers and the traffic flow that this allows [31]. Traffic flow is an open system, and it develops waves of traffic jams (solitons) excited by the density of traffic. Transitions between different traffic flow patterns have been considered as phase transitions, typical products of self-organization in the non-living world.

A number of examples of self-organizing systems from different fields is given in Table 1, lower section.

Explanatory Concepts of Self-Organization

Despite half a century of inquiry, the theory of self-organizing systems is still in its infancy. There is no "standard model" of SOS, only various aspects emphasized by different researchers. Here we shall discuss the most important of these.

Non-Equilibrium Thermodynamics

Thermodynamics has been concerned with the notion of order and disorder in physical systems for more than a century. The theory of self-organization has to address fundamental issues of this field. The most important question in this regard is, how order can arise through self-organization.

Classical thermodynamics has focused on closed systems, i. e. systems isolated from external influence in the form of matter and energy flow. This allowed to understand the processes involved when a system evolves undisturbed. A key result of this inquiry is the second law of thermodynamics, originally formulated by Carnot and later refined by Clausius in the 19th century. It states that "any physical or chemical process under way in a system will always degrade the energy". Clausius introduced a quantitative measure of this irreversibility by defining entropy:

$$S \equiv \int \mathrm{d}Q/T \tag{1}$$

with Q the heat energy at a given temperature T. In any process of a closed system, entropy always rises

$$\frac{\mathrm{d}S}{\mathrm{d}t} \geq 0\,. \tag{2}$$

According to Eddington, 1928 [32] this universal increase in entropy "draws the arrow of time" in nature.

Boltzmann had reformulated entropy earlier in terms of the energy microstates of matter. In his notion, entropy is a measure of the number of different combinations of microstates in order to form a specific macrostate.

$$S = k_{\mathrm{B}} \ln(W) \tag{3}$$

with k_{B} Boltzmann's constant and W the thermodynamic probability of a macrostate. He argued that the macrostate with most microstates (with maximum entropy) would be most probable and would therefore develop in a closed system. This is the central tenet of equilibrium thermodynamics.

More interesting phenomena occur if the restrictions for isolation of a system are removed. Nicolis and Prigogine [11] have examined these systems of non-equilibrium thermodynamics which allow energy and matter to flow across their boundary. Under those conditions, total entropy can be split into two terms, one characterizing internal processes of the system, $\mathrm{d}_{\mathrm{i}}S$ and one characterizing entropy flux across the border $\mathrm{d}_{\mathrm{e}}S$. In a generalization of the second law of thermodynamics, Prigogine and Nicolis postulated the validity of the second law for the internal processes,

$$\frac{\mathrm{d}_{\mathrm{i}}S}{\mathrm{d}t} \geq 0 \tag{4}$$

but explicitly emphasized that nothing can be said about the sign of the entropy flux. Specifically, it could carry a negative sign and it could be larger in size than the internal entropy production. Since the total entropy is the sum of both parts, the sign of the total entropy change of an open system could be negative,

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \frac{\mathrm{d}_{\mathrm{i}}S}{\mathrm{d}t} + \frac{\mathrm{d}_{\mathrm{e}}S}{\mathrm{d}t} < 0 \tag{5}$$

a situation impossible in equilibrium thermodynamics. Thus, increasing order of the system considered would be possible through export of entropy. Self-organization of a system, i. e. the increase of order, would not contradict the second law of thermodynamics. Specifically, the non-equilibrium status of the system could be considered a source of order.

Even in the distance from thermodynamic equilibrium, however, certain stable states will occur, the *stationary* states. These states assume the form of *dissipative structures* if the system is far enough from thermodynamic equilibrium and dominated by non-linear interactions. The preconditions for dissipative structures can be formulated as follows:

1. The system is open.
2. The inner dynamics is mainly non-linear.
3. There are cooperative microscopic processes.
4. A sufficient distance from equilibrium is assumed, e. g. through flows exceeding critical parameter values.
5. Appropriate fluctuations appear.

If those conditions are fullfilled, the classical thermodynamic branch of stationary solutions becomes unstable and dissipative structures become stable system solutions.

Synergetics

Prigogine's description of dissipative structures is formally limited to the neighborhood of equilibrium states. As

Haken pointed out, this is a severe restriction on its application and in particular precludes its formal application to living systems. Instead, Haken proposed *order parameters* and the *slaving principle* as key concepts for systems far from equilibrium. Let the time evolution of a continuous dynamical system is described by

$$\frac{\mathrm{d}\mathbf{q}}{\mathrm{d}t} = \mathbf{N}(\mathbf{q}, \alpha) + \mathbf{F}(t) \tag{6}$$

where $\mathbf{q}(t) = [q_1(t), q_2(t), \ldots, q_N(t)]$ is the system's state vector and $\mathbf{N}$ is the deterministic part of the system's interaction whereas $\mathbf{F}$ represent fluctuating forces, and α are the so-called control parameters. Then the stable and unstable parts of the solution can be separated by linear stability analysis, as can the time dependent and time independent parts. As a result, the solution can be written as

$$\mathbf{q}(t) = \mathbf{q}_0 + \sum_u \xi_u(t)\mathbf{v}_u + \sum_s \xi_s(t)\mathbf{v}_s \tag{7}$$

$\mathbf{v}_u, \mathbf{v}_s$ are the unstable and stable modes, respectively, and $\xi_u(t), \xi_s(t)$ are their amplitudes. These amplitudes obey the following equations

$$\frac{\mathrm{d}\xi_u}{\mathrm{d}t} = \lambda_u \xi_u + N_u(\xi_u, \xi_s) + \mathbf{F}_u(t) \tag{8}$$

$$\frac{\mathrm{d}\xi_s}{\mathrm{d}t} = \lambda_s \xi_s + N_s(\xi_u, \xi_s) + \mathbf{F}_s(t) \tag{9}$$

with λ_u, λ_s characterizing the linear part of the equations and function N summarizing the non-linear deterministic components. The slaving principle formulated by Haken now allows to eliminate the stable mode development by expressing them as a function of unstable modes

$$\xi_s(t) = f_s[\xi_u(t), t] \,. \tag{10}$$

Thus, the unstable modes (order parameters) enslave the stable modes and determine the development of the system's dynamics. This result is useful both to describe *phase transitions* and *pattern formation* in systems far from equilibrium.

Synergetic concepts have been applied in a variety of disciplines [33].

Chaos and Complexity

The treatment of *chaotic systems* has been derived from non-linear system theory. Chaotic systems are usually low-dimensional systems which are unpredictable, despite being deterministic. The phenomenon was originally discovered by the meteorologist E. Lorenz in 1963 [34], although H. Poincare in 1909 was aware of the possibility of certain systems to be sensitive to initial conditions [35]. The reason for the difficulty to predict their behavior stems from the fact that initially infinitesimal differences in trajectories can be amplified by non-linear interactions in the system. These instabilities, together with the lack of methods for solving even one-dimensional non-linear equations analytically, produce the difficulties for predictions. Modern theory of deterministic chaos came into being with the publication of a seminal article by May in 1976 [36].

Complex systems, on the other hand, have many degrees of freedom, mostly interacting in complicated ways, i. e. they are high-dimensional. All the more astonishing is the fact that our world is not totally chaotic in the sense that nothing can be predicted with any degree of certainty. It became apparent, that chaotic behavior is but one of the ways non-linear dynamical systems behave, with other modes being complex attractors of a different kind.

Complexity itself can be measured, notably there exist a number of complexity measures in computer science, but describing or measuring complexity is not enough to understand complex systems.

Self-Organized Criticality

For particular high-dimensional systems, Bak et al. [43] have suggested a dynamic system approach toward the formation of fractal structures, which are found to be widespread both in natural and artificial environments. Their canonical example was a pile of sand. They examined the size and frequency of avalanches under certain well-prepared conditions, notably that grains of sand would fall on the pile one by one. This is an open system with the forces of gravity and friction acting on the possibly small fluctuations that are caused by deviations in the hitting position of each grain of sand. They observed how the grains would increase the slope of the sand pile until more or less catastrophic avalanches developed.

Bak suggested the notion of *self-organized criticality* (SOC) as a key concept which states that large dissipative systems drive themselves to a critical state with a wide range of length and time scales. This idea provided a unifying framework for the large-scale behavior in systems with many degrees of freedom. It has been applied to a diverse set of phenomena, e. g. in economic dynamics and biological evolution. SOC serves as an explanation for many power-law distributions observed in natural, social and technical systems, like earthquakes, forest fires, evolutionary extinction events, and wars. As opposed to the widely studied low-dimensional chaotic systems, SOC systems have a large number of degrees of freedom, and still exhibit fractal structures as are found in the extended space-time systems in nature.

The Hypercycle

In a series of contributions since 1971, Eigen and Schuster have discussed particular chemical reaction systems responsible for the origin, self-organization and evolution of life [37,38,39,40]. By considering *autocatalytic sets* of reactions they arrived at the most simple form of organization, the *hypercycle*, which is able to explain certain aspects of the origin of life. They have considered a chemical reaction system composed of a variety of self-reproductive macromolecules and energy-rich monomers required to synthesize those macromolecules. The system is open and maintained in a non-equilibrium state by a continuous flux of energy-rich momomers. Under further assumptions they succeeded in deriving Darwinian selection processes at the molecular level. Eigen and Schuster have proposed rate equations to describe the system.

The simplest system realizing the above mentioned conditions can be described by the following rate equations

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = (A_i Q_i - D_i)x_i + \sum_{k \neq i} w_{ik} x_k + \Phi_i(\mathbf{x}) \qquad (11)$$

where i enumerates the individual self-reproducing units and x_i measures their respective concentrations. Metabolism is quantified by the formation and decomposition terms $A_i Q_i x_i$ and $D_i x_i$. The ability of the self-reproducing entities to mutate into each other is summarized by the quality factor for reproduction, Q_i, and the term $w_{ik} x_k$ which takes into account all catalytic productions of one sort using the other. A_i, D_i are rate constants for selfreproduction and decay respectively. The flow term Φ_i finally balances the production/destruction in this open system in order to achieve $\sum_k x_k = \text{const.}$

By introducing a new feature called excess production

$$E_i \equiv A_i - D_i \qquad (12)$$

and its weighted average

$$\bar{E}(t) = \sum_k E_k x_k / \sum_k x_k \qquad (13)$$

and symbolizing the "intrinsic selection value" of a sort i by

$$W_{ii} = A_i Q_i - D_i \qquad (14)$$

one arrives at reduced rate equations

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = (W_{ii} - \bar{E})x_i + \sum_{k \neq i} w_{ik} x_k \,. \qquad (15)$$

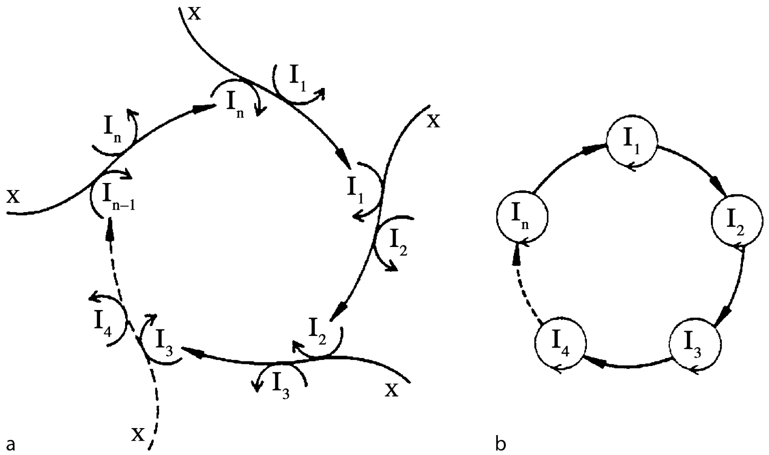
These equations can be solved under certain simplifying assumptions and notably yield the concept of a *quasi-species* and the following extremum principle: A quasi-species y_i is a transformed self-replicating entity with the feature that it can be considered as a cloud of sorts x_i whose average or consensus sequence it is. The extremum principle reads: Darwinian selection in the system of quasi-species will favor that quasi-species which possesses the largest eigen-value of the rate equation system above.

The Origin of Order

In the 1990s Kauffman [41] pointed out one of the weaknesses of Darwinian theory of evolution by natural selection: It cannot explain the '*origin* of species' but rather only their subsequent development. Kauffman instead emphasized the tendency of nature to constrain developments along certain paths, due to restrictions in the type of interaction and the constraints of limited resources available to evolution. In particular he held up the view that processes of spontaneous order formation conspire with the Darwinian selection process to create the diversity and richness of life on Earth.

Previously, Kauffman had formulated and extensively studied [42] the NK fitness landscapes formed by random networks of N Boolean logic elements with K inputs each. Kauffman observed the existence of cyclic attractor states whose emergence depended on the relation between N and K, and the absolute value of K. In the case of large K ($K \approx N$), the landscape is very rugged and behavior of the network appears stochastic. The state sequence is sensitive to minimal disturbances and to slight changes of the network. The attractor length is very large, $\approx N/2$, and there are many attractors. In the other extremal case, $K = 2$, the network is not very sensitive to disturbances. Changes of the network do not have strong and important consequences for the behavior of the system.

Kauffman proposed NK networks as a model of regulatory systems of living cells. He further developed the notion of a *canalizing function* that is a Boolean function in which at least one variable in at least one state can completely determine the output of the function. He proposed that canalizing functions are an essential part of regulatory genetic networks.



Self-organizing Systems, Figure 1
The Hypercycle (reproduced from [15])

Emergence and Top-Down Causation

The notion of *emergence* has been introduced in complex systems theory in order to explain the appearance of new qualitative features on the level of an entire system that could not be observed at the level of its components. Emergent behavior can be connected to the afore-mentioned complex attractors. It requires switching the level of description of behavior of a system, from local (component-centered) to global (system-centered), or at least to a meso-level (sub-system-centered). Emergent behavior happens when

a) the system shows qualitatively new behavior on a higher level of description which
b) could not have been easily predicted from the interactions of components at the lower level (obeys a non-linear relationship)
c) is the result of a self-organization process.

Emergence is strongly related to self-organization. It is often understood as a pattern formation process. While it essentially has to do with changing the perspective and looking at the system at a different level, it concerns itself with a change in behavior (e. g. the system is getting more organized, shows new coordinated modes of behavior). It has been further conjectured that there is top-down causation, i. e. the structures forming on the higher level of the system are able to affect the lower levels (system components) and influence them in a way that stabilizes the newly emergent behavior. Haken could show in the context of Synergetics that this phenomenon exists. *Top-down causation* is believed to be an important source of complexity, especially in living systems, because it stabilizes patterns.

Self-organization draws heavily from this source of qualitative innovation in complex systems.

Modeling Methods

A formal model is a simplified mathematical or algorithmic representation of a system. Often it has been simplified to the point of a carricature, and this has to be born in mind when making conclusions about the consequences of model predictions. No model can predict beyond the limits of its approximations.

Mean-Field Methods

One of the most important methods used to model complex systems is tied to the notion of dynamical systems. Dynamical systems are systems whose time development is accessible to a description by state changes. It entails the existence of a state space in which these changes can be traced and quantified.

Mean-field methods of description focus on average behavior. They abstract away from the local correlations between a system's elements and describe only long-range changes. For instance, the behavior of a planet could be described as a point on its trajectory around the star it circles. Detailed interactions of its atmosphere would not be part of that description.

Mean-field methods are formulated in the form of time-dependent differential or difference equations which can be solved under certain conditions and predict the behavior of a system in its state space.

Assuming that the state of a system can be subsumed in a vector of state variables **x** whose values are observable and depend on time, we can generally formulate an equation for continuous time development as:

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}) . \tag{16}$$

If time does not develop continuously, a discrete (iterative) equation can be used to describe system behavior:

$$\mathbf{x}_{t+1} = \mathbf{f}(\mathbf{x}_t) . \tag{17}$$

The notion is that the state of the system at time t, $\mathbf{x}(t)$ or $\mathbf{x}_t$, is everything that is of interest and can be known about the system. It turned out that with this extreme simplification many systems became treatable that would have otherwise resisted quantitative treatment. The non-linear nature of many of the state development equations, however, and the high-dimensionality of state space vectors often constitute prohibitive hurdles to exact or even approximate mathematical solution of these equations.

As a result, algorithmic approaches for modeling self-organizing systems have become more prevalent in recent years.

Agent-Based Models

A very general class of algorithmic systems is subsumed under the term agent-based models. In these systems, individual entities are modeled that interact with each other. Thus, the approximation of average behavior, and the interest for long-term behavior only is abandoned in favor of a microscopic description of the elements of a system and their interactions. The abstraction of features of a system is achieved through the assumption of rules of behavior of the agents, including their interaction behavior. Agent-based systems must be implemented as computational systems, and run on a computer to obtain results. Agents are assigned states, and transition rules between states, depending on interacting agents, and then these rules are executed in parallel over the set of agents under consideration.

Cellular Automata A particular subset of agent-based models is the class of cellular automata introduced by von Neumann [47] going back to lattice networks of Ulam. The agents of cellular automata are placed on a grid of cells and allowed to assume a finite number of states. Interactions are determined by state transition rules and the definition of a neighborhood, which determines the interactivity of the cellular automaton. Many variants of cellular automata exist, differing in the number of dimensions of the grid, the number of states, the sort and distribution of transition rules and the nature of the neighborhood.

A typical cellular automata model might, e. g. consist of digital cells (allowing only two states, "ON" and "OFF"), homogeneous and deterministic transition rules between states, a one-dimensional grid, and nearest-neighbor interactions. Cellular automata of this type have been thoroughly examined in [44,45] and show a surprising variety and richness in behavior.

In a cellular automaton like LIFE, for instance, one can observe how macroscopic and mesocopic structures appear through self-organization, that is, as a process determined solely by the local interaction of the CA's elements. some structures, e. g. spiral waves, are more resilient against perturbation than others, e. g. glider canons. A moving structure like a glider can be interpreted as an emergent phenomenon as it does not seem to be present on the microscopic scale (single CA cells do not move).

Graphs and Networks A more general class of automata can be formulated if the notion of cellular neighborhood is abandoned. Instead of a rigidly defined grid, a graph or network of automata connected through edges to other automata is introduced. Each node of the graph/network represents an automaton, with interactions allowed via edges.

The notion of a graph is, however, more general, and allows other agent-based systems to be simulated. For instance, the nodes of a graph might represent species of an ecosystem interacting with other species (connected by edges). Each species might be represented by a state counting the number of individuals of that species. Nodes might further hold information on particular features of individuals, and possibly their variants. Explicit simulations of such systems have been considered in the context of "Artificial Chemistries" [46].

In recent years, the structure and dynamics of networks has been a major focus of interest in the scientific community. Network science has become a converging point for different disciplines interested in modelling complex behavior.

Observables

Self-organizing phenomena rest on the appearance of particular sets of behaviors. If ever they are to be understood, a clear notion of observable quantities needs to be developed that allows a proper description of the behavior of such systems. At present, no such canonical set of observables exists, owing to the bewildering variety of systems that show signs of self-organization. However, one can discern a number of different measures and observables that might form the core of such a set [49].

Entropic and Information Theoretic Measures One class of observables can be considered entropic and information-theoretic measures. These measures have in common a statistical root, and seek to describe a self-organizing system in terms of the order (or disorder) that develops over time [48].

Stability Measures Another class of observables can be discussed as stability measures. In this class, systems are sought to be disturbed from their regular behavior in order to obtain a clearer idea of their resilience.

Scaling Measures A further class of observables can be attached to features of scaling. Both theoretical and experimental approaches can be used to vary the number of dimensions, number of equations/agents, number and complexity of interactions, etc, in these systems. Scaling behavior can then be observed for particular quantities and systems classified accordingly.

Patterns and Flows The defining observables of a self-organizing system are patterns. These refer to the collective behavior of the elements of a system, differentiating them from noise. If individual entities would not show such correlations in their behavior, self-organization could not be observed. Patterns can be described in a variety of ways, e. g. as multidimensional vectors, using spatial and temporal coordinates. If patterns change dynamically one can speak of flows.

The central tenet of self-organization is that systems exist whose pattern forming tendencies are determined by themselves, and not by an outside agency.

The Role of Self-Organization in Science, the Social Sciences and Engineering

Self-organization as a concept has assumed center stage in Science. With the advent of nonlinear systems and studies on complex systems in non-equilibrium situations, the explanatory power of self-organization now permeates every branch of scientific enquiry.

From structure formation at the level of super-galactic clusters, even starting from the development of the entire universe, down to microscopic particles and their interaction patterns, self-organizing phenomena have been postulated, theorized, observed and confirmed.

In particular the origin and evolution of life have been studied under the aspect of self-organization. Within Biology, the developmental process of organisms as well as their metabolisms, growth and learning have been identified as self-organizing processes.

In the humanities, the idea of self-organization has taken roots, although the paradigm is far from being fully recognized yet. Since the 1990s the origin and development of languages has been an object of study under the premise of self-organization. In social science the concept of self-organization has been studied since a number of years, due to the obvious fact that interaction between social actors generate a society. Even in psychology, self-organizing principles begin to appear.

Economy and Management Science have taken notice of the concept, and a growing number of enterprise concepts promote the idea of a form as a self-organizing entity.

Finally, Philosophy has embraced the concept of self-organization and connected it to earlier thoughts on the development of the scientific process and epistemology. Whitehead put forward his process philosophy, and Smuts, already in the 1920s, promoted the notion of holism which has strong connections to self-organization. Evolutionary epistemology was formulated as a response to traditional epistemology and emphasizes the aspect of natural selection affecting senses and cognitive abilities.

Engineering is beginning to grasp the ubiquity of self-organization in Nature. Specifically in the area of nanotechnology the concept is used extensively for the purpose of self-assembly of molecular entities. At nanoscales, it is very difficult to directly specify the structuring behavior of entities. As a result, self-organizing properties of matter are used to the advantage of the structural outcome.

Different kinds of infrastructure networks have been recognized as self-organizing, and Engineering begins to make use of the tendency of networked systems to self-organize.

In the area of adaptation, there exists a long tradition of making use of self-organization principles. The self-organizing feature map, introduced by Kohohen, has been a key step forward in the domain of unsupervised learning of artificial neural networks.

Open Issues and Future Directions

So far, there is no unique theory of self-organization. Over the course of many years different approaches have been used, but a coherent picture has not yet emerged.

An important open question in the area of the mathematical basis for self-organization is the formulation of a theory of *constructive (evolutionary) systems*, that is systems which, in the course of their development, generate new elements that subsequently interact with elements already created earlier.

Another question aims at the raison d'etre of *hierarchical systems*. Why do they form, how do they structure themselves, and what would be possible to apply from these principles in Engineering? Notably, how would one build self-organizing systems such that they do something useful? How could they be controlled?

In Science, the build-up of complexity remains a controversial issue. Is it true that evolution of the universe tends to increase complexity, or is there no tendency of complexity increase at all? What are the mechanisms by which Nature increases complexity, if any? How could we apply this knowledge in planning and managing complexity in the human world?

A wealth of questions remains, and it is anticipated that the 21st century will shed light on at least a few of them.

Cross References

- ▶ Entropy in Ergodic Theory
- ▶ Mathematical Basis of Cellular Automata, Introduction to
- ▶ Swarm Intelligence
- ▶ Biological Development and Evolution, Complexity and Self-organization in
- ▶ Rough and Rough-Fuzzy Sets in Design of Information Systems

Bibliography

Primary Literature

1. von Bertalanffy L (1968) General System Theory: Essays on its Foundation and Development. George Braziller, New York
2. Wiener N (1965) Cybernetics: Or Control and Communication in the Animal and the Machine. MIT Press, Cambridge
3. Ashby WR (1947) Principles of the Self-Organizing Dynamic System. J Gen Psychol 37:125–128
4. Farley BG, Clark WA (1954) Simulation of self-organizing systems by digital computer. IRE Trans Inf Theor 4:76–84
5. Ashby WR (1960) Design for a Brain: The Origin of Adaptive Behavior. Chapman and Hall, London
6. von Foerster H (1960) On Self-organizing systems and their environments, In: Yovits MC, Cameron S (eds) Self-Organizing Systems. Pergamon Press, New York, pp 31–50
7. Haken H (1983) Synergetics – An Introduction, 3rd edn. Springer, Berlin
8. Kauffman S (1993) The Origins of Order. Oxford University Press, Oxford
9. Conrad M (1983) Adaptability. Plenum Press, New York
10. Kohonen T (2000) Self-Organizing Maps. Springer Series in Information Sciences, vol 30. Springer, Berlin
11. Nicolis G, Prigogine I (1977) Self-Organization in Nonequilibrium Systems. Wiley, New York
12. Haken H (2000) Information and Self-Organization: A Macroscopic Approach to Complex Systems. Springer Series in Synergetics, 2nd edn. Springer, Berlin
13. Thom R (1989) Structural Stability and Morphogenesis: An Outline of a General Theory of Models. Addison-Wesley, Reading
14. Mandelbrot B (1982) The fractal Geometry of Nature. W.H. Freeman, San Francisco
15. Eigen M, Schuster P (1979) The Hypercycle. Springer, Berlin
16. Rosen R (1991) Life Itself: A Comprehensive Inquiry into the Nature, Origin, and Fabrication of Life. Columbia University Press, New York
17. von der Malsburg C (1973) Self-organization of orientation sensitive cells in the striate cortex. Kybernetik 14:85–100
18. Linsker R (1988) Self-organization in a perceptual network. Computer 21:105–117
19. von Foerster H (2002) Understanding Understanding. Springer, New York
20. Maturana H, Varela F (1979) Autopoiesis and Cognition. Reidel, Dordrecht
21. Camazine S, Deneubourg J, Franks N, Sneyd J, Bonabeau E, Theraulaz G (2000) Self-Organization in Biological Systems. Princeton University Press, Princeton
22. Sole R, Bascompte J (2006) Self-Organization in Complex Ecosystems. Princeton University Press, Princeton
23. Luhmann N (1995) Social Systems. Stanford University Press, Palo Alto
24. Focardi S, Cincotti S, Marchesi M (2002) Self-organization and market crashes. J Econ Behav Organ 49:241–267
25. Portugali J (2000) Self-Organization and the City. Springer, Berlin
26. Bruckner S, di Marzo Serugendo G, Karageorgos A, Nagpal R (eds) (2005) Engineering Self-Organizing Systems. Springer, Berlin
27. Lovelock JE, Margulis L (1974) Atmospheric homoeostasis by and for the biosphere: The Gaia hypothesis. Tellus 26:2–9
28. Lenton TM (1998) Gaia and natural selection. Nature 394:439–447
29. Staley M (2002) Darwinian Selection Leads to Gaia. J Theor Biol 218:35–46
30. Kerner BS (1998) Experimental Features of Self-Organization in Traffic Flow. Phys Rev Lett 81:3797–3800
31. Treiber M, Helbing D (1999) Explanation of Observed Features of Self-Organization in Traffic Flow. Arxiv preprint cond-mat/9901239
32. Eddington AS (1928) The Nature of the Physical World. Gifford Lectures. Cambridge University Press, Cambridge
33. Haken H (2004) Synergetics. Introduction and Advanced Topics. Springer, Berlin
34. Lorenz E (1963) Deterministic Nonperiodic Flow. J Atmospheric Sci 20:130–141
35. Poincare H (1909) Science et Methode. Flammarion, Paris
36. May RM (1976) Simple Mathematical Models with very complicated Dynamics. Nature 261:459–467
37. Eigen M (1971) Selforganization of Matter and the Evolution of Biological Macromolecules. Naturwissenschaften 58:465–523
38. Eigen M, Schuster P (1977) The hypercycle. A principle of natural self-organization. Part A: Emergence of the hypercycle. Naturwissenschaften 64:541–565
39. Eigen M, Schuster P (1978) The hypercycle: a principle of natural self-organization, Part B. Naturwissenschaften 65:7–41
40. Eigen M, Schuster P (1978) The hypercycle: a principle of natural self-organization, Part C. Naturwissenschaften 65:341–369
41. Kauffman SA (1993) The Origins of Order. Oxford University Press, Oxford
42. Kauffman SA (1969) Metabilic Stability and epigenesis in randomly constructed nets. J Theor Biol 22:437–467

43. Bak P, Tang C, Wiesenfeld K (1988) Self-organized criticality. Phys Rev A 38:364–374
44. Wolfram S (1983) Statistical Physics of Cellular Automata. Rev Mod Phys 55:601–644
45. Wolfram S (1984) Cellular Automata as Models of Complexity. Nature 311:419–424
46. Dittrich P, Ziegler J, Banzhaf W (2001) Artificial Chemistries – A Review. Artif Life 7:225–275
47. von Neumann J (1966) Theory of Self-Reproducing Automata. Univ of Illinois Press, Chicago
48. Polani D (2003) Measuring Self-Organization via Observers, In: Banzhaf W et al (eds) Advances in Artificial Life. Lecture Notes in Artificial Intelligence, vol 2801. Springer, Berlin, pp 667–675
49. Shalizi CR, Shalizi KL, Haslinger R (2004) Quantifying Self-Organization with Optimal Predictors. Phys Rev Lett 93:118701

Books and Reviews

Bak P (1996) How Nature Works – The Science of Self-Organized Criticality. Springer, New York

Bar-Yam Y (2003) Dynamics of Complex Systems. Westview Press/ Perseus Books, New York

Boccara N (2004) Modelling Complex Systems. Springer, New York

Edelman G (1987) Neural Darwinism. Basic Books, New York

Hemelrijk C (ed) (2005) Self-organization and Evolution of Social Systems. Cambridge University Press, Cambridge

Holland J (2000) Emergence: From Chaos to Order. Oxford University Press, Oxford

Jantsch E (1980) The Self Organizing Universe: Scientific and Human Implications. Pergamon Press, New York

Johnson S (2001) Emergence. Scribner, New York

Kelso S (1995) Dynamic Patterns: The Self-organization of Brain and Behavior. MIT Press, Cambridge

Lovelock JE (1995) The Ages of Gaia. W.W. Norton, New York

Mingers J (1995) Self-Producing Systems. Plenum Press, New York

Morowitz H (2002) The Emergence of Everything. Oxford University Press, Oxford

Rosen R (1999) Essays on Life Itself. Columbia University Press, New York

Sornette D (2003) Critical Phenomena in Natural Sciences. Springer, Berlin

Self-replicating Robotic Systems

JACKRIT SUTHAKORN[1], MATTHEW MOSES[2], GREGORY S. CHIRIKJIAN[2]

[1] Biomedical and Robotics Technology Lab, Faculty of Engineering, Mahidol University, Salaya, Thailand

[2] Robot and Protein Kinematics Lab, Department of Mechanical Engineering, Johns Hopkins University, Baltimore, USA

Article Outline

Glossary

Artificial life Artificial life refers to the study of artificial systems that exhibit lifelike properties, typically involving some form of self-replication and/or evolution. This is a very broad field of study and includes work involving computer simulations, cellular automata, chemistry, robotics, and synthetic biology.

Cellular automata A cellular automaton (CA) is a theoretical construct where a collection of cells are organized into regular grids or lattices. Many arrangements are possible, but typically one-dimensional CAs are composed of square cells arranged in a line, and two-dimensional CAs are composed of square cells arranged in a square grid. Each cell contains a finite state machine. The state of all cells in a CA are typically updated synchronously (at the same time) with each cell changing to a new state that is a function of its previous state and the previous states of its neighbors. The surrounding cells that affect a given cell's state transition are called the cell's neighborhood. CAs can simulate a wide variety of physical processes by designing an appropriate neighborhood and state transition function. CAs are most often studied by implementing them with computer simulations.

Finite state machine A finite state machine (FSM) is a conceptual computing machine with an input, output, and memory. At regular intervals in time the machine transitions (changes) to a new state, which is a function of the current state and the machine's current input. The output is a function of the input and the state. The function of the memory is to store the state information between each transition. FSMs are easily implemented with discrete digital electronic components, microcontrollers, or computer simulations. Implementations of FSMs are widely used to control industrial devices and consumer electronics.

Modular robot A modular robot is composed of distict "modules" which contain motors, mechanisms, electronics, and interconnections. Modules are typically designed so that they can easily be connected and disconnected from each other. In some modular robotic systems, many identical modules are assembled into larger robots. This allows the same set of modules to form robots optimized for different tasks. In other cases, the modules may be specialized for certain tasks, and when assembled they form a robot with additional functionality. Motivation for building modular robots includes increased versatility, ease of replacing damaged components, and potentially lower manufacturing costs.

Self-reconfigurable robot A self-replicating robotic system is a robotic device that exhibits some form of self-replication. This chapter is concerned primarily with directed robotic self-replication, in which a robotic device interprets some form of coded instructions in order to carry out the replication process. In nature deoxyribonucleic acid (DNA) typically encodes replication instructions. In the examples of robotic self-replication presented in this chapter, the instructions may be encoded in a computer program, an arrangement of modular components, or as a pattern of lines that guide the motion of a mobile robot.

Self-replication Self-replication is the process by which an entity creates a duplicate of itself. The most familiar example is the self-replication of living organisms, although other natural processes such as crystal growth can be classified as self-replication.

Universal constructor A universal constructor (UC) is a conceptual machine that reads instructions and executes them to construct an object. A key property of a UC is that it can construct any object which can be described to it via the instructions, including duplicates of itself. UCs have been demonstrated in computer simulations using cellular automata. Modern manufacturing tools, such as assembly robots and computer-controlled machine tools, are similar to UCs but practical limitations make these machines "somewhat less than universal" constructors. The ribosome, a complex molecule present in nearly all biological cells, performs a function very similar to that of a UC, assembling proteins according to instructions encoded in messenger ribonucleic acid (mRNA).

von Neumann universal constructor The mathematician John von Neumann proposed a cellular automata model of a universal constructor capable of self-replication. Many researchers have refined and improved von Neumann's original design since it was first presented in the 1950s. The original design used tens of thousands of cells in a two-dimensional CA with 29 states. The general structure of the model is a movable constructing arm controlled by instructions encoded in a long line of cells resembling a tape, conceptually similar to a computer-controlled robot arm.

Introduction

The concept of an artificial self-replicating system was introduced in the 1950s by John von Neumann [31]. Von Neumann introduced the theory of self-replicating automata and established a quantitative definition of self-replication. His early results on self-replicating machines have become useful in several diverse research areas such as: cellular automata, nanotechnology, macromolecular chemistry, and computing [8,23,26,27]. However, prior to the turn of the millennium, a fully autonomous self-replicating physical robot had never been implemented. In this chapter, a series of prototype designs from our laboratory and their physical implementation are described. We begin by discussing some motivation and history, then go on to describe a remote-controlled replicating robotic system and a semi-autonomous replicating robotic system. We then describe some fully autonomous self-replicating systems, and discuss how manufacturing work cells might be designed so as to reproduce.

Motivation

People have imagined for years a factory that could autonomously replicate itself for multiple generations, requiring neither people nor the monstrous machinery typically associated with a factory. Over recent decades, outer space has been mentioned as one potential application for such self-replicating robotic factories (see e. g., [3,10], and [7]). However, enormous technical barriers must be overcome before these systems can become feasible. The purpose of the current work is to take one small step toward realizing this goal.

In contrast to self-reconfigurable robotics [1,12,14,21,34], self-replication utilizes an original unit to actively assemble an exact copy of itself from passive components. This has the potential to result in exponential growth in the number of robots available to perform a job, thus drastically shortening the original unit's task time.

Descriptions of Self-Replication by Johns von Neumann

According to von Neumann there are four components required for a self-replicating machine: the builder, the controller, the copier, and the blueprint [31]. The process of self-reproduction begins with the controller commanding the builder to fabricate an exact mechanical system by following the blueprint. Then the controller instructs the copier to replicate the blueprint, input the new blueprint in the replica, and start the new machine. The process described in the previous paragraph was written as a mathematical model called the "Von Neumann Kinematic Beast" by Rolf Pfeifer et al. of the University of Zurich [24], which goes as follows:

1. Let A represent the builder. If a machine G is desired, one can build this machine with the blueprint, *Blue(G)*.

   $$A + \mathit{Blue}(G) \rightarrow G\,.$$

   Where '+' indicates that the machine is composed of the left and right components ($A + \mathit{Blue}(G)$) and '→' indicates construction.
2. Let B represent the copier. B would make a copy of the blue print, *Blue(G)*.

   $$B + \mathit{Blue}(G) {\rightarrow} \mathit{Blue}(G)\,.$$

3. Let C be the controller. With the combination of A and B, this would trigger them to follow in the correct order to generate a desired mechanical system G and the new blueprint *Blue(G)*, and then wrap them up together and split them from the original machine

   $$A + B + C + \mathit{Blue}(G) \rightarrow G + \mathit{Blue}(G)\,.$$

4. Let G be the machine $A + B + C$ then we get:

$$A{+}B{+}C{+}\mathit{Blue}(A{+}B{+}C) \rightarrow A{+}B{+}C{+}\mathit{Blue}(A{+}B{+}C).$$

Any system that observes this is a self-reproducing machine in Neumann's view. A short time after successfully presenting some of his theoretical work, von Neumann began working on the implementation of his theories. In the 1950's, with the help of Stanislaw Ulam, von Neumann invented the Cellular Automata concept (redrawn and shown in Fig. 1.)

Previous Efforts in Mechanical Self-Replicating System

Von Neumann [31] was the first to seriously study the idea of self-replicating machines from a theoretical perspective. In the late 1950's, Penrose performed the first recognized demonstration of a self-replicating mechanical system [22]. It consisted of passive elements that self-assembled under external agitation. This is similar in many ways to the modern work of Whitesides [32], only at a different length scale. Moore [19] was interested in von Neumann's concepts, but he commented that Neumann's Cellular Automata was only for demonstration, and that applications of self-replicating systems needed to be carried out. Moore described several conceptual designs of artificial living plants which could duplicate themselves not only from off-the-shelf artificial parts, but also from materials from nature. Jacobson [13] constructed a self-replicating machine using parts of toy trains. His replication was done on a round section of toy track.

In the 1980s, NASA established a series of studies on the topic of "Advanced Automation for Space Missions" [10]. These studies investigated the possibility of building a self-replicating factory on the moon. References [5,6,7,9,30] also outlined strategies for space utilization. Recently, research on robots that are capable of designing other machines with little help from humans has also been performed (see [17] and references therein). This uses rapid prototyping technologies.

Our Previous Replicating Prototypes

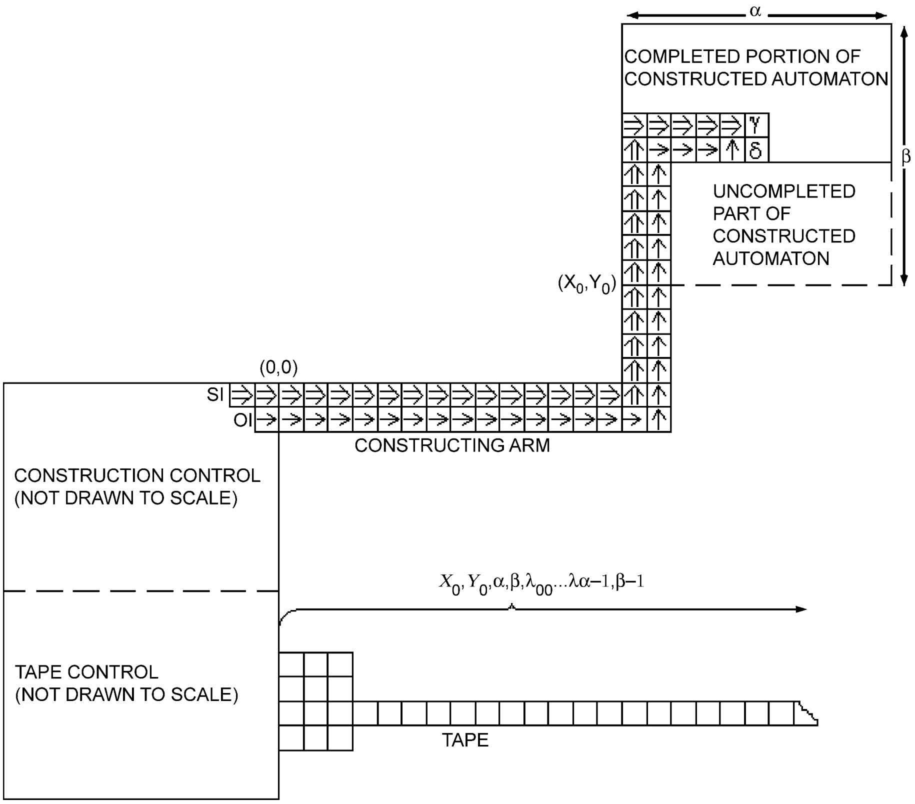
This section describes in chronological order the robotic prototypes capable of various levels of self-replication that have been designed and built in our laboratory. Our prototypes are constructed from modified LEGO Mindstorm kits with enhanced electrical connections because of their modularity, functionality and ease of use.

Demonstration I: Prototype 1

We built this robot as the first prototype to demonstrate that it is mechanically feasible for one simple robot to produce a copy of itself. This robot depends on external passive fixtures for self-replication. Figure 2 shows fixtures, the original prototype 1, and a set of the replica's subsystems. (Details of Prototype 1 can be found in [27].)

Demonstration II: Prototypes from the Spring 2002 Mechatronics Class at JHU.

From experience gained from Demonstration 1, we introduced the concept of self-replicating robots to students in a hands-on Mechatronics course (taught in the Department of Mechanical Engineering at the Johns Hopkins University) under the supervision of the last author, and TA'ed by the first author. We divided students in this course into eight groups to explore designs and implementations of the concept of self-replicating robotic systems. In order to focus on the mechanical issues involved in the design of the self-replicating systems, the robots


α
β
COMPLETED PORTION OF CONSTRUCTED AUTOMATON
γ
δ
UNCOMPLETED PART OF CONSTRUCTED AUTOMATON
(X0,Y0)
(0,0)
SI
OI
CONSTRUCTING ARM
CONSTRUCTION CONTROL (NOT DRAWN TO SCALE)
X0,Y0,α,β,λ00...λα–1,β–1
TAPE CONTROL (NOT DRAWN TO SCALE)
TAPE

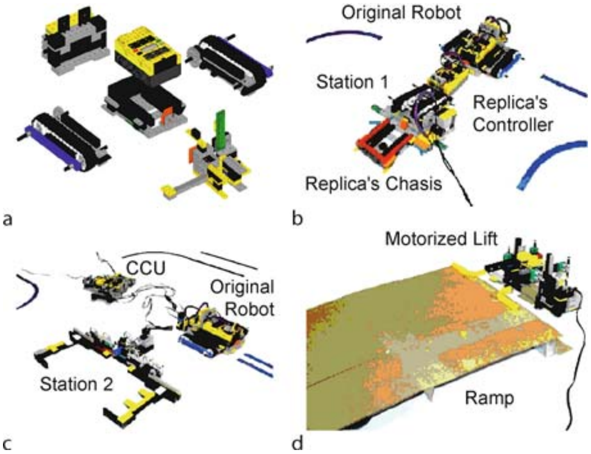
Self-replicating Robotic Systems, Figure 1
Schematic of von Neumann's self-reproducing cellular automaton (Redrawn from [31])

Self-replicating Robotic Systems, Figure 2
Fixtures, the original prototype 1, and a set of the replica's subsystems

were remote-controlled rather than autonomous. (Details of Demonstration II can be found in [27], and [2].)

Demonstration III: A Semi-Autonomous Replicating System

This work builds upon previous results in remote-controlled robotic replication with a new feature that many subtasks in the replication process are now autonomously performed by the robot. The use of feedback sensors was implemented. The robot is unable to directly build



Self-replicating Robotic Systems, Figure 3
a An assembly view of the semi-autonomous robotic system, b Assembling station 1, c Assembling station 2, and d Assembly station 3

its replicas. Therefore, several work-cells (intermediate robots) are required to assist the original robot in the replicating process. Each work-cell works as a station in this factory-liked replicating system. Figures 3a, b, c, and d illustrate an assembly view of the original robot, stations 1, 2, and 3, respectively. Full details of the semi-replicating robot can be found in [27,28].

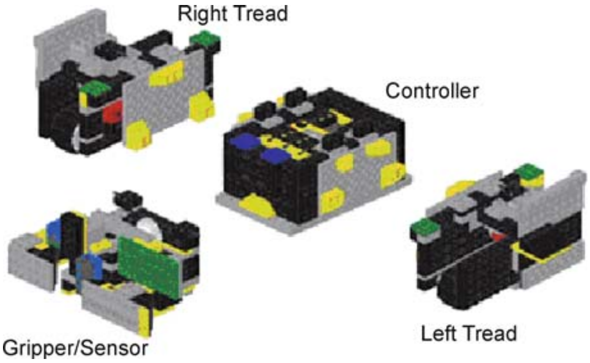
Design and Descriptions of an Autonomous Self-Replicating Robot

The robot and its replicas each consist of four subsystems: controller, left tread, right tread, and gripper/sensor subsystems. All subsystems are connected to others using magnets and shape constraints. Figure 4 shows an assembly view of the robot.

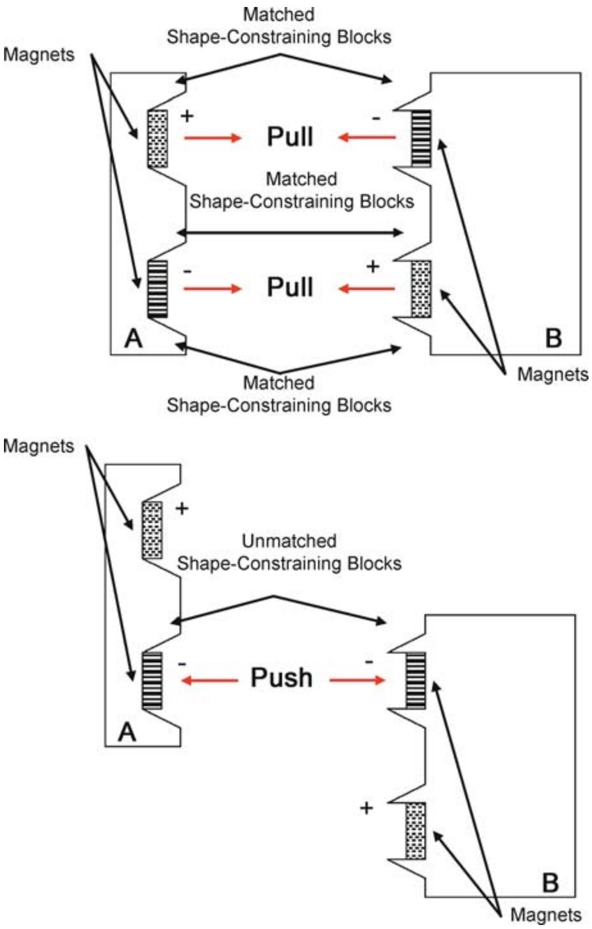
The controller subsystem is made up of a LEGO RCX programmable controller fit inside a chassis. The chassis's sides are used to connect to the left and right treads. Each side has a set of magnets, a set of shape-constraining blocks, and a set of electrical connections. The front end of the chassis is designed to attach with the gripper. The front end also has a set of magnets, a set of shape-constraining blocks, and a set of electrical connections, which transfer electrical signals and power from the controller to the gripper's motor and the navigating sensors installed on the gripper subsystem.

The magnets and the shape-constraining blocks are used in collaboration to aid aligning and interlocking subsystems. On each chassis side, the magnets are symmetrically placed in opposite polar directions to each other. This is to protect against incorrect positioning of the subsystems. The concept of using the magnets (with different polarizations) and shape-constraining blocks was influenced by the self-complementary molecules of Rebek [25]. Figure 5 illustrates the concepts of using the polar magnets and shape-constrained blocks to align and interlock subsystems. By design, it is very difficult for these connectors to misalign.

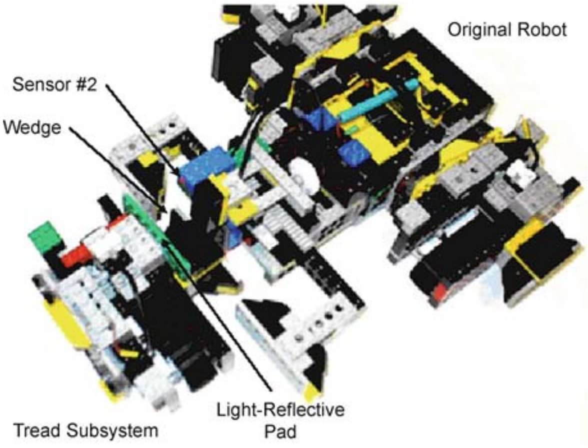
The left and right tread subsystems are designed to be identical to each other, with the purpose of reducing the system's design complexity. A tread subsystem hosts a rubber tread with a driving gear system, a 9 V LEGO DC motor, and a light-reflective pad which helps the original robot's navigation. One side of the tread has a set of magnets, a set of shape-constrained blocks, and a set of electrical connections, all of which correspond to the side of the controller subsystem. On the other side, the tread has a wedge which is fitted to the gripper. The wedge is used during the tread subsystem's transferring and assembling processes. Figure 6 shows how the original robot grasps



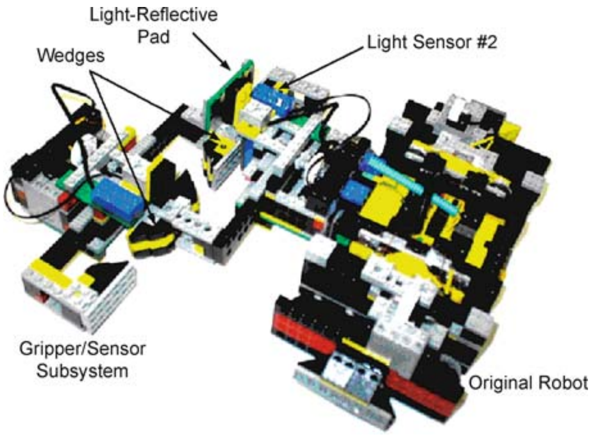
Self-replicating Robotic Systems, Figure 4
An assembly view of the self-replicating robot


Matched
Shape-Constraining Blocks
Magnets
+
-
Pull
Matched
Shape-Constraining Blocks
-
+
Pull
A
B
Magnets
Matched
Shape-Constraining Blocks
Magnets
+
Unmatched
Shape-Constraining Blocks
-
Push
-
+
A
B
Magnets

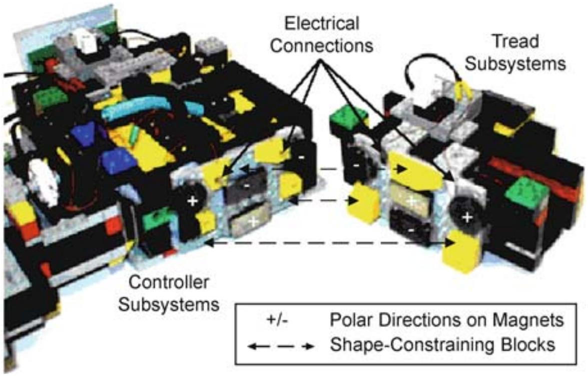
Self-replicating Robotic Systems, Figure 5
This diagram illustrates the concept of using polar magnets and shape-constraining blocks (*top*: correctly aligning, and *bottom*: incorrectly aligning)



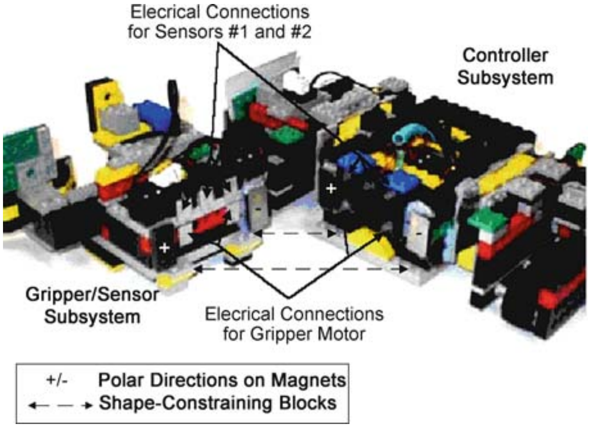
Self-replicating Robotic Systems, Figure 6
The original robot grasps the tread subsystem



Self-replicating Robotic Systems, Figure 8
The original robot grasps the gripper/sensor subsystem


Electrical Connections
Tread Subsystems
Controller Subsystems
+/- Polar Directions on Magnets
←– – → Shape-Constraining Blocks

Self-replicating Robotic Systems, Figure 7
The connections located between controller and tread subsystems


Elecrical Connections for Sensors #1 and #2
Controller Subsystem
Gripper/Sensor Subsystem
Elecrical Connections for Gripper Motor
+/- Polar Directions on Magnets
←– – → Shape-Constraining Blocks

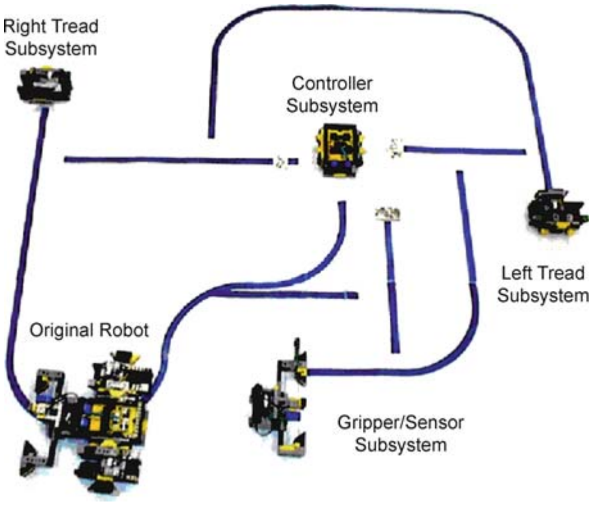
Self-replicating Robotic Systems, Figure 9
The connections located between gripper/sensor and controller subsystems

the tread subsystem, and Fig. 7 shows the connections located between the controller and tread subsystems.

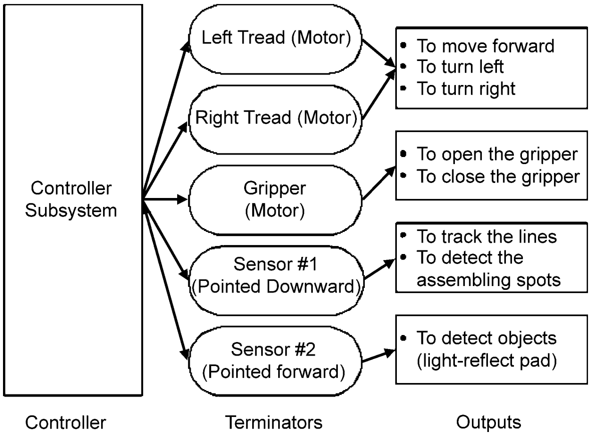
The gripper/sensor subsystem is comprised of a 9 V LEGO DC motor, a set of rack and pinion gears used to drive the left/right fingers of the gripper, a set of magnets, a set of shape-constrained blocks, a set of electrical connections, and two light sensors (one is pointed downward, and the other is pointed forward).

The set of magnets, shape-constraining blocks, and electrical connections are attached to their corresponding part, on the front side of the controller subsystem. The left finger of the gripper is designed in a wedge shape to be fitted with the gripper in any identical robot. This wedge is used in the same manner as in the tread subsystems during assembling processes. Figure 8 shows how the original robot grasps the gripper/sensor subsystem, and Fig. 9 shows the connections located between gripper/sensor and controller subsystems. The two LEGO light sensors are employed in the robot's navigation system. The first light sensor (pointed downward) is used to detect the blue painted lines and silver acrylic spots on the experiment surface. The second light sensor (pointed forward) is used to detect objects (the subsystems of the replica) which the robot runs into.

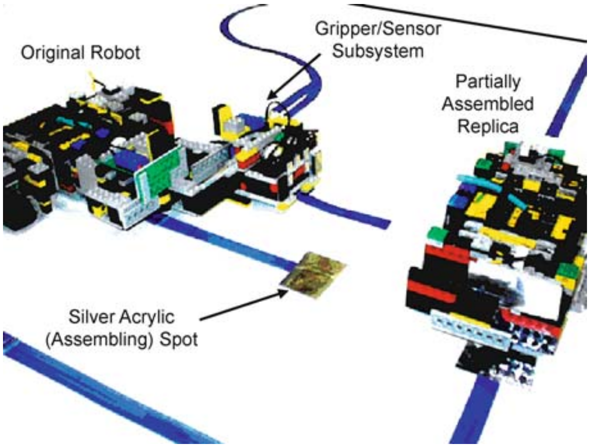
The experimental area is a 2 m x 3 m area made of white colored paper with lines and spots painted in blue and silver acrylic colors. The original robot starts at the initial position, and the replica's subsystems are at their locations. Figure 10 shows the experimental area with locations of the replica's subsystems and the initial position of the original robot.



Self-replicating Robotic Systems, Figure 10
A map of the experimental area



Self-replicating Robotic Systems, Figure 11
The control architecture of the autonomous self-replicating robot



Self-replicating Robotic Systems, Figure 12
The robot is searching for the assembly location while holding the replica's gripper/sensor subsystem

Controls and Programming of the Autonomous Self-Replicating Robot

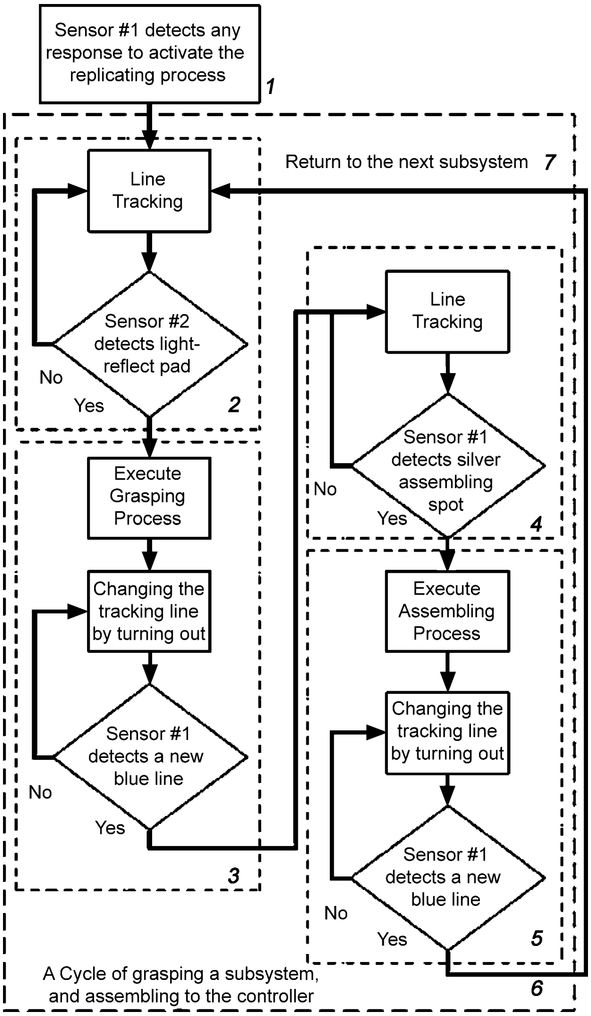
The prototype robot is a fully autonomous system. Figure 11 shows the conceptual control architecture of the robot. The robot and its replicas, using the LEGO light sensor No. 1 (pointed downward), is capable of tracking the blue lines, and it can recognize the assembling spots, painted in a silver acrylic. The sensor detects and returns different analog values, corresponding to different colors. The robot tracks the painted lines to navigate between positions. Once the robot detects the assembling spot, the robot begins the assembling process. The LEGO light sensor No. 2 (pointed forward) returns an analog value once it detects a light-reflective pad attached to the tread and gripper/sensor subsystems. This notifies the robot to begin grasping the detected subsystem.

The grasping process consists of an aligning push toward the subsystem, and closing the gripper to grasp the subsystem. On the other hand, the assembly process consists of opening the gripper to release the subsystem, and an aligning push forward to snap the subsystem to the controller. Figure 12 shows the original robot grasping a subsystem, and moving toward an assembling spot, in silver acrylic, along the blue line.

The programming of the prototype is described here. The code is programmed on a PC and transferred through a LEGO infrared program-transferring tower. In the order in which events take place in the replication process, the programming is separated into seven stages: 1) replication process is activated, 2) line tracking and searching for a subsystem, 3) grasping the subsystem and changing to a new path which leads to the next step, 4) line tracking and searching for the assembly location, 5) assembling the subsystem to the controller and changing to a new path, 6) the new path leads to the next subsystem, 7) The final step loops back to steps 2 through 6 so the process repeats indefinitely. Figure 13 illustrates the programming flowchart of the self-replicating robot system.

Experiments and Results of the Autonomous Self-Replicating Robot

The following is a step-by-step outline of the procedure that our autonomous self-replicating robot system under-


Sensor #1 detects any response to activate the replicating process
1
Return to the next subsystem 7
Line Tracking
Sensor #2 detects light-reflect pad
No
Yes
2
Execute Grasping Process
Changing the tracking line by turning out
Sensor #1 detects a new blue line
No
Yes
3
Line Tracking
Sensor #1 detects silver assembling spot
No
Yes
4
Execute Assembling Process
Changing the tracking line by turning out
Sensor #1 detects a new blue line
No
Yes
5
A Cycle of grasping a subsystem, and assembling to the controller
6

Self-replicating Robotic Systems, Figure 13
The flowchart of the self-replication process

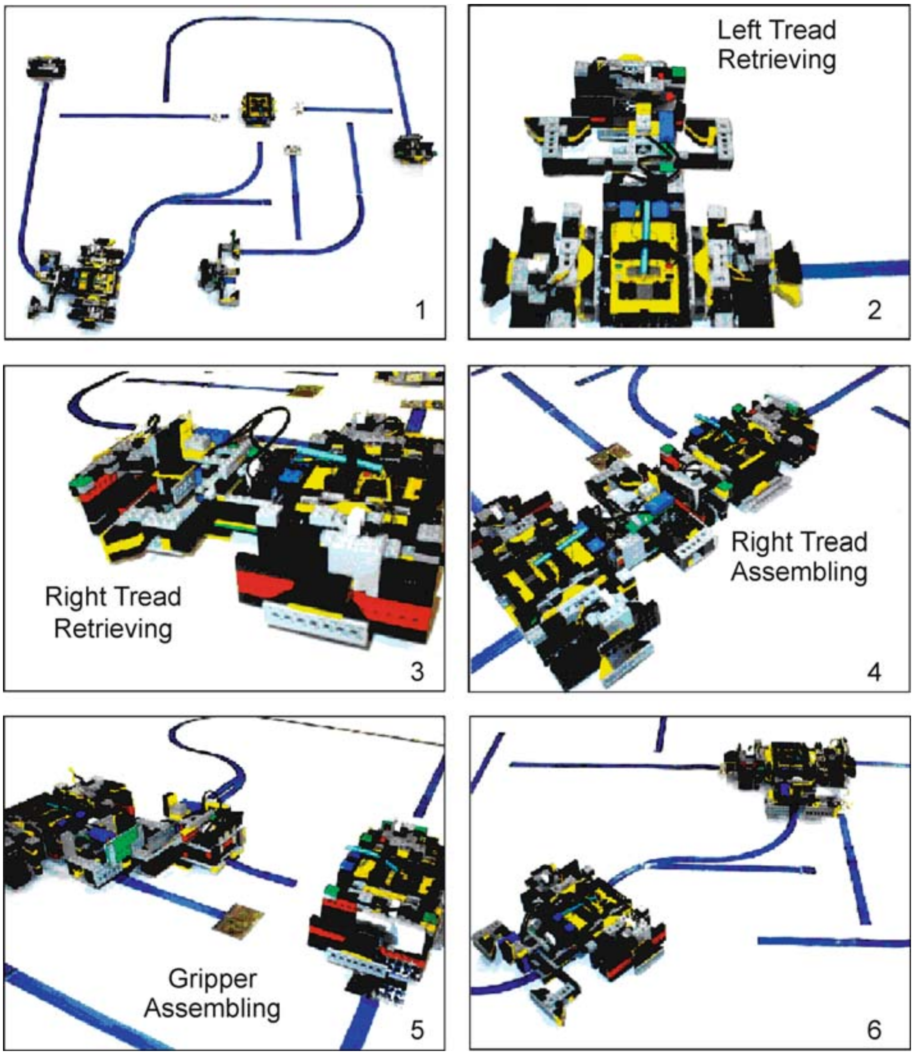
goes (Fig. 14 is a photographic representation of these step but they are not synchronous):

1. The original robot starts following the line from the starting point to the first subsystem using sensor No. 1.
2. Once light sensor No. 2 detects the first subsystem (right tread), the original robot begins the grasping process and grasps the right tread subsystem.
3. After the grippers are closed the original robot turns to the right until it detects a line.
4. The original robot follows the second line until it reaches the assembly location.
5. When light sensor No. 1 on the original robot detects the silver acrylic spot (the assembly location), the robot stops, and begins the attaching process.
6. The original robot opens the grippers, and gives a final push to secure the right tread subsystem to the controller subsystem.
7. The original robot then backs up and turns to the left until it detects a line value on sensor No. 1.
8. The original robot follows the line until it reaches the left tread subsystem.
9. Once light sensor No. 2 detects the second subsystem, the robot will stop, and begin the grasping process by closing its gripper around the left tread's wedge.
10. The original robot turns right until it detects the next line.
11. The original robot will follow the second line until it reaches the assembly location.
12. The original robot opens its gripper to release the left tread subsystem.
13. The original robot gives a final push on the left tread subsystem to help secure it.
14. The original robot then backs up and turns left until it detects the next line, using sensor No. 1.
15. The original robot follows the line to the final subsystem.
16. Once it reaches the gripper/sensor subsystem, it stops, and begins the grasping process.
17. The original robot closes its gripper, and turns right until it detects a line value with sensor No. 1.
18. The gripper/sensor subsystem is now transferred to the assembly location.
19. Once the original robot reaches the assembly location it stops, and opens the gripper.
20. The original robot backs up and turns left until sensor No. 1 is a line value.
21. The original robot then follows the line back to the starting point, and is ready to replicate again.
22. The completed replica self-activates (20 seconds after completion) and begins following the line to the starting point.
23. Once each robot reaches the starting point, it begins the replication procedure again.

The replication process takes two minutes and fifteen seconds per cycle. Although each subsystem is required to be placed in its starting location, errors in initial position and orientation are not very critical. We found slight errors during the grasping process in a few experiments caused by improper placement of the subsystems. Overall, the system is robust and very repeatable.

Our Approach to Self-Replicating Control Circuitry

Ideally, for a robotic system to be truly self-replicating, it would have to demonstrate the ability to assemble all of



Self-replicating Robotic Systems, Figure 14

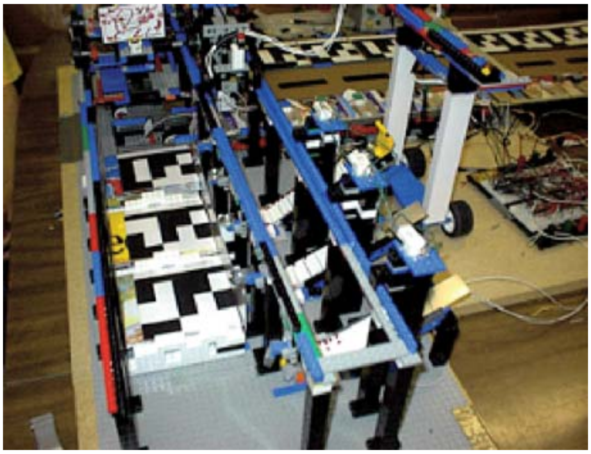
Self-replication process: 1) the original robot begins at the initial position with every part placed in their position. 2) The robot detects the right-tread subsystem. 3) The robot grasps the right-tread subsystem, and attaches it to the controller. 4) In the same manner, the robot performs the assembly of the left-tread subsystem. 5) After the robot grasps the gripper/sensor subsystem the robot transfers the subsystem to the next assembly step at the controller. 6) After being fully assembled, the replica is self-activated, and ready to replicate just like the original

its own subsystems from the most fundamental components. In the case of the robot controller, we consider the most fundamental components to be transistors, resistors, capacitors, etc., whereas microcontrollers are too complex to be considered as basic elements.

Our approach is to build a circuit capable of controlling an electro-mechanical system to re-build replicas of the control circuit from the most fundamental electronic components. In the von Neumann universal constructor paradigm, an associated instruction code is also required. In contrast it is possible to replicate a particular system by self-inspection without invoking von Neumann's universal constructor. We illustrate both concepts in hardware designed and constructed by students in a Mechatronics course taught at Johns Hopkins University in 2003. Two prototypes illustrate replication by self-inspection, and one demonstrates the universal constructor. In all three cases, pre-built electro-mechanical systems (called the SRI-builders) use the transistorized circuit as its controller. While in the von Neumann paradigm, the controller follows instructions that are explicitly encoded (and hence must reproduce the code for the overall system to be self-replicating), in the self-inspection paradigm, actions are taken implicitly as a result of observing the spatial layout of components in the original and feeding that information into the circuit itself. Clever electromechanical design ensures that observations obtained during self-inspection are translated directly into actions without requiring the interpretive step of consulting a long sequence of encoded construction commands.

A von Neumann Universal Constructor Prototype

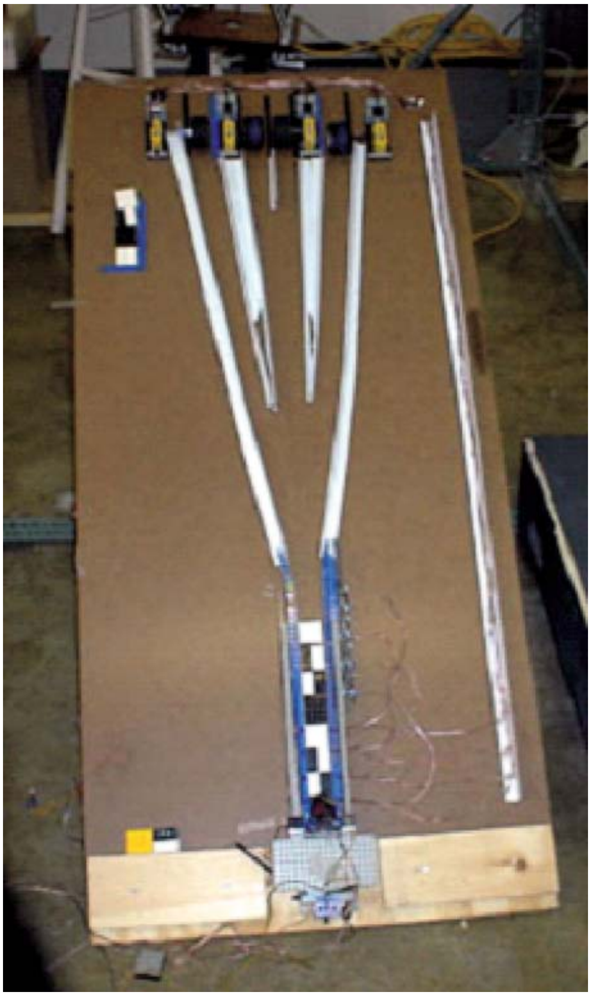
The prototype is a two-arm gantry-style robot with two degrees of freedom. The first degree of freedom moves along the whole system, includes the feeders and the assembly boards. The second degree of freedom moves vertically to pick and place codes and circuit pieces. A controller circuit is used to control motions of the robot. There are two boards being assembled at a certain amount of time. The first board is the replicated circuit, and the second is the replicated codes. The circuit is pre-wired before it is placed into the carrier. This is similar to a process of placing a chip into a circuit board. In the code part, each code consists of three lines of black and white strip representing 3 bits. A series of codes is set up on a code array where it is fed simultaneously to a reader array. The reader array is made of an array of photo-transistors and infrared LED emitters used to read black (0) and white (1) colors in each code. Infrared LED emitters are used instead of regular LED for reducing problem with ambient light distractions. Figure 15 shows the system in a side view. Once the acquisition part of the system reads the code, the assembly part of the system works by following the instruction code, replicating the code and circuit of that part.



Self-replicating Robotic Systems, Figure 15
Side view of the replicating system (von Neumann universal constructor prototype)

Non-Universal Self-Replication by Self-Inspection (Design 1)

This design is the first design of a non-universal self-replication by self-inspection. The self-replicating control circuit has the ability to identify the proper electronic components required, translate information about its own constituent parts obtained from self-inspection into mechanical tasks that create a replica, and transfer all functions to the replica. There is no list of instructions in the form of a code. Each electronic component has a black-and-white color code. Parts are loaded into feeders, and as a reading head traverses the control circuit, the information about which part of the control circuit is being observed is fed into the circuit itself. This actuates the solenoid in the ap-



Self-replicating Robotic Systems, Figure 16
Side view of the replicating system (non-universal self-replication by self-inspection – design 1. (See [11] for details))

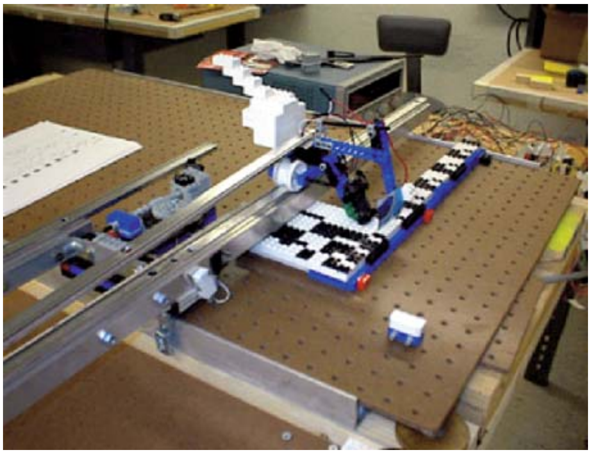
propriate feeder to release the parts needed to form the replica. Parts then slide down an incline and form an orderly array. The reading head continues to move and creates replicas until resources are completely utilized or its track ends. The design is scalable and the components are modular, allowing many different levels of intelligence to be replicated. This concept is one of many which we are investigating to enable self-replicating robots to perform complex behaviors. See Fig. 16.

Non-Universal Self-Replication by Self-Inspection (Design 2)

This robotic system is an X-Y table constructed from modified LEGO components. A photo-transistor sensor system is attached to the end-effector of the X-Y system in order to inspect the control circuit (the components of which are each assigned a unique black and white code). On the top of the X-Y system, a set of component feeders is installed. The circuit converts the signal from the sensor system to control the component feeders to release the correct component to the parts assembler. The parts assembler then arranges all the components to create a new replica of the control circuit. See Fig. 17.

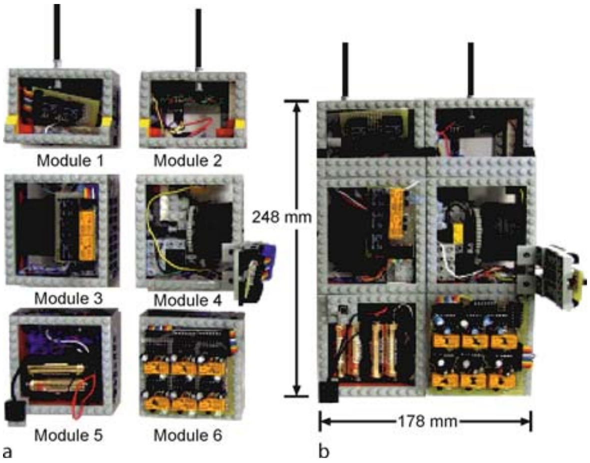
Merging the Self-Replicating Robot and Self-Replicating Circuit Concepts

In prior prototypes we considered two disjoint cases: (1) self-replicating robots constructed from modules where one module contained a computer; (2) a control circuit that commanded a mechanical device to assemble copies of the circuit. In this section we review very recent work with K. Lee and the JHU authors that merges these two concepts.

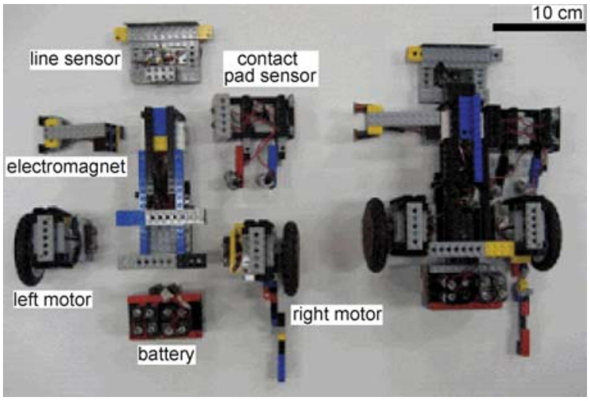


Self-replicating Robotic Systems, Figure 17
Side view of the replicating system (non-universal self-replication by self-inspection – design 2)

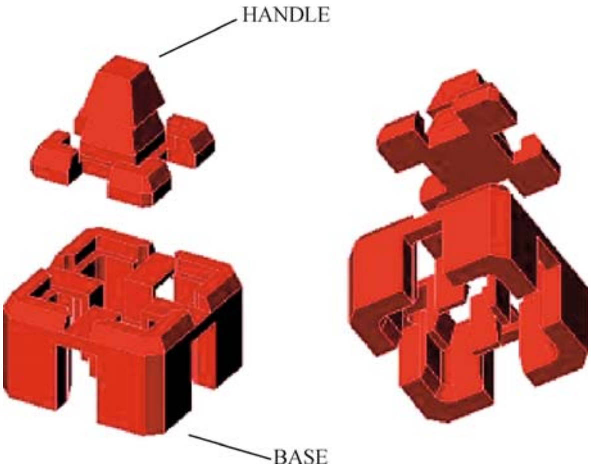
Basically, the autonomous self-replicating robot described previously in this paper functions as a finite-state machine in a highly structured environment. This same behavior can be implemented by a robot controlled by a simple circuit without using a microprocessor. Discrete electronics elements of this circuit such as transistors, resistors, etc., can be distributed over the modules from which the robot is constructed. This means that as the original robot assembles these modules to form a replica, the controller for the replica is assembled as the robot is. Prototypes are shown in Figs. 18 and 19.



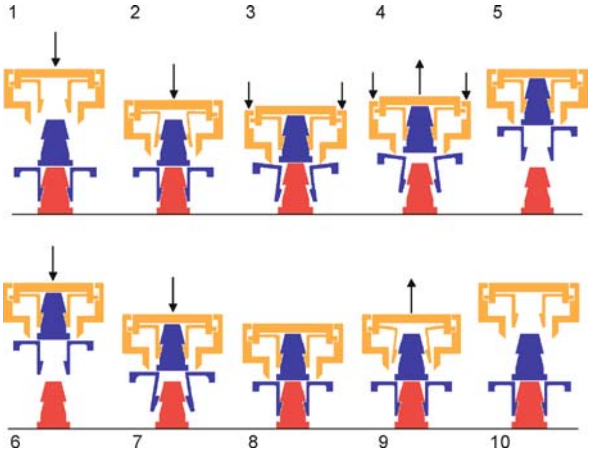
Self-replicating Robotic Systems, Figure 18
A modular self-replicating robot controlled by a finite state circuit (see [15,16])



Self-replicating Robotic Systems, Figure 19
Example of another modular self-replicating robot controlled by a finite state machine (see [4,16,18] for details)



Self-replicating Robotic Systems, Figure 20
Two exploded views of the basic component. The handle and base are individually cast in polyurethane, then bonded together with epoxy



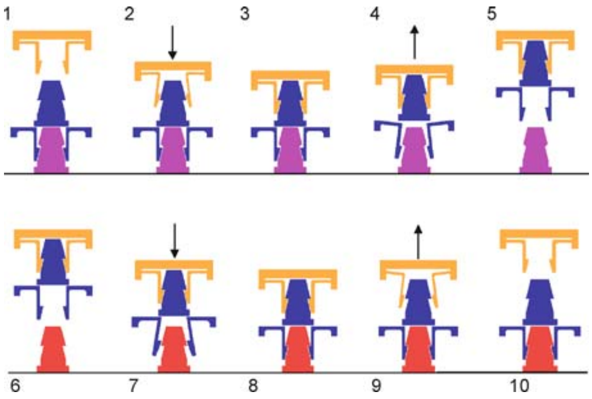
Self-replicating Robotic Systems, Figure 21
Steps in reversible assembly process. The tool (top part) grasps, retrieves, and reconnects one component to another

Towards a Universal Constructor

Whereas other prototypes of self-replicating robotic systems use relatively few subsystems as the initial parts, and connect these subsystems with magnets, the desire to create machines that can reproduce from a large number of basic parts requires thought about mechanical design, manufacturing, and assembly issues. Therefore, in this section the work of the second author in [20] is reviewed. In this work, a novel set of mechanical components that can be assembled into a wide variety of devices is presented. The components are specifically designed to be handled and assembled by devices made of the same type of components. A 3-axis Cartesian manipulator built with these components is presented. It is shown that in principle the manipulator can assemble duplicates of itself in addition to arbitrary devices when provided with properly oriented components and controlled by a human operator.

All components in the component set are variants of the "basic component". Figure 20 shows a cutaway view of two basic components. Each component consists of two parts: an upper "handle" and a lower "base". The base dimensions are 3.8 by 3.8 by 2.3 cm. The handle and base are individually cast from polyurethane resin in a silicone mold and then fixed together with epoxy adhesive. The base contains four compliant snap tangs that are complementary to the four tapered surfaces of the handle. Figure 21 shows how the snap tangs of the base grasp the undercut of the handle of another part. Additionally, the convex tapered surfaces of the base mate with concave tapered surfaces of the handle in order to provide rigidity with respect to shear and torsion between parts. Chamfers on the edges of the parts allow a certain amount of positioning error during assembly. The tangs and handle also contribute to error tolerance, since they are tapered such that two parts need not be exactly positioned before assembly. The slots cut in each side of the base allow the part to be assembled onto other parts that have reinforcing segments. The tangs are L-shaped, and the top part of each tang is exposed by a slot cut in each side of the handle.

The parts have a common handle, so they all can be manipulated by a single grasping tool. A part can be connected to another part by simply lowering it into place such that the tangs engage the handle of the other part (sliding components are slid sideways onto the handle of the other part). Depending on the grasping tool used, part connections can be reversible or irreversible. A section view of the steps in reversible assembly are shown in Fig. 21. The grasping tool (upper part) has thin tangs and a release mechanism. The process is initiated when the grasping tool is lowered onto the part to be disconnected. The release mechanism is engaged, depressing the tangs of the part to be disconnected. The upper part of the tangs deflect, disengaging the tangs of the middle part from the handle of the lower part (step 3). In step 4 the grasping tool and middle part are lifted from the lower part. To re-connect the part, the grasping tool and part are lowered onto the destination part (step 7). The grasping tool is removed simply by lifting it from the newly connected part. The snap tangs in the grasping tool are thinner than normal tangs, so they deflect first and disengage the grasping tangs from the handle on the newly connected part (step 9).

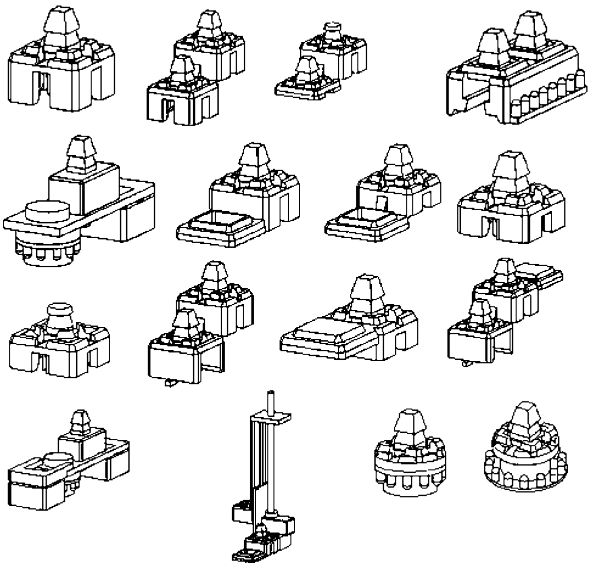


Self-replicating Robotic Systems, Figure 22
Steps in irreversible assembly process. The tool retrieves a part from a special storage site and attaches it to another component

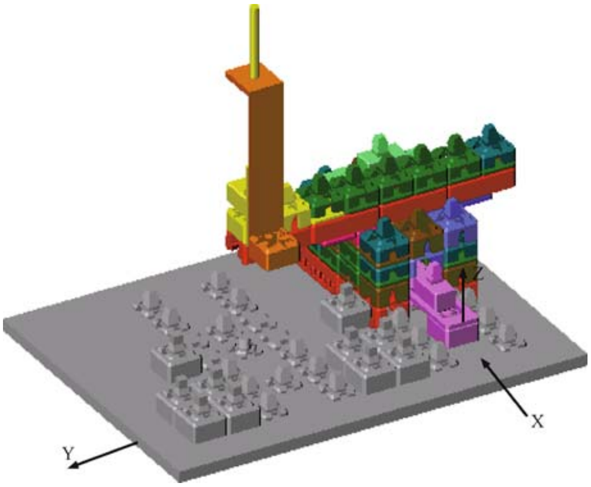
Irreversible assembly is also possible, as shown in Fig. 22. In this case, the grasping tool has thin tangs like the reversible tool, but there is no release mechanism. Parts must be stored at a special "storage site" before assembly. The storage site (lower part in steps 1–5) is a normal part with reduced undercut on the handles. The amount of undercut on the storage site handles and the thickness of the tangs on the tool are carefully chosen so that the force required to disassemble two components is greatest between two normal parts, and least between a normal part and the storage site. This allows the grasping tool to pick a part from a storage site and connect it to another part with two simple up/down vertical movements. The irreversible assembly method, due to its simplicity, is used in the manipulator described below.

These basic components can be used to form a constructing machine capable of assembling a wide variety of devices, including duplicates of itself. By "capable" is meant kinematically capable - the machine can grasp, manipulate, and place all of the components required for constructing a duplicate of its "mechanical self". The machine's "mechanical self" includes the hardware required to manipulate mechanical components, but excludes the portion responsible for control. The machine lacks any type of sensor or onboard control, and so must be controlled either by a human operator or a sophisticated external controller. By "assemble" is meant the retrieval of components from a storage site and subsequent connection of them to an assembly or operating plane. Components must arrive at the storage site through action of another entity – namely a human operator.

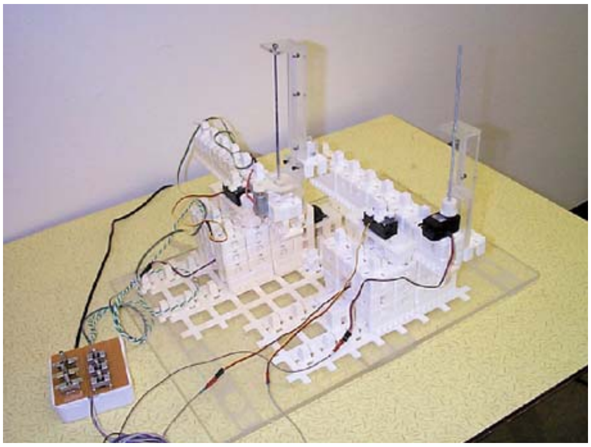
A total of 16 different variations on the basic component were built. Most of these are used in the manipulator, and the remainder are useful for constructing other types of devices. Figure 23 shows a diagram of these parts. Figure 24 shows a CAD model of a constructing machine made from parts in the component set. It is essentially a 3-axis Cartesian manipulator. The base of the machine is a platform that slides along the x-axis, and is actuated by a DC motor. A motor-driven boom rides on top of the platform, and slides along the y-axis. The end-effector is mounted to a vertically-linear motor on the end of the boom. The reachable space of the end-effector is a rectangular volume of size 3 × 4 × 6 in units of component-widths. Note that the tool does not rotate. This requires



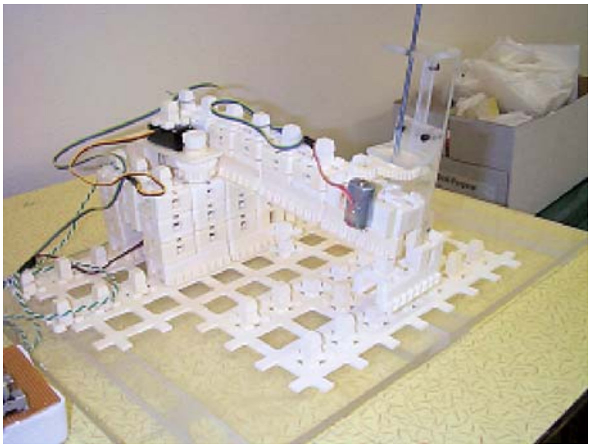
Self-replicating Robotic Systems, Figure 23
Sixteen variations on the basic component



Self-replicating Robotic Systems, Figure 24
A constructing machine made from parts in the component set



Self-replicating Robotic Systems, Figure 25
An initial physical implementation of a universal constructor [20]



Self-replicating Robotic Systems, Figure 26
Large downward deflection of extended boom

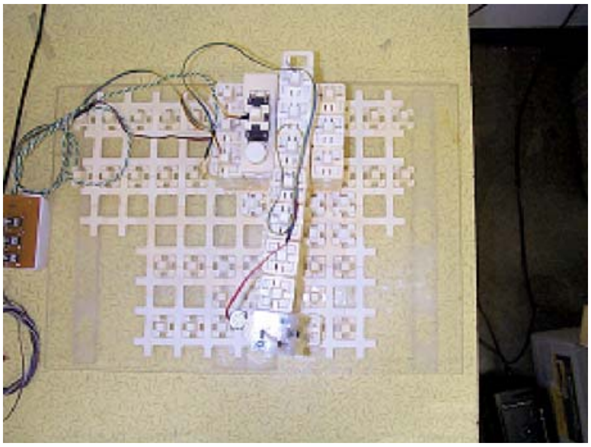
parts to be loaded at the storage site in their proper orientation. The constructor has a total of 45 components of eleven different types, not counting the operating plane. The details of the assembly process of this machine can be found in [20].

Figure 25 shows the experimental setup in its finished state. Both devices are operable. A human operator controls them through a switchbox that turns the various motors on or off. A topic that has been neglected so far is that of running wires to the motors. As seen from the figure, this is done rather haphazardly. This prototype had a number of problems that are worth mentioning which are helping us to design better systems.

The first problems that were encountered dealt with getting the constructor boom to move smoothly. When the boom is extended, it applies high forces to the parts that hold it to the platform. This results in more friction, and requires a higher torque from the motor. In these conditions the motor tended to separate from the platform, and the drive gear would lose contact with the racks on the boom. This problem was eventually "solved" by using a metal screw and adhesive between certain parts to hold them in place.

Most of the other problems occurred during assembly, and centered around the extended boom. Figure 26 shows the large downward deflection occurring in the extended boom. This made it difficult to align parts. A human operator could successfully place parts, using repeated effort and visual feedback, but a simple controller surely could not.

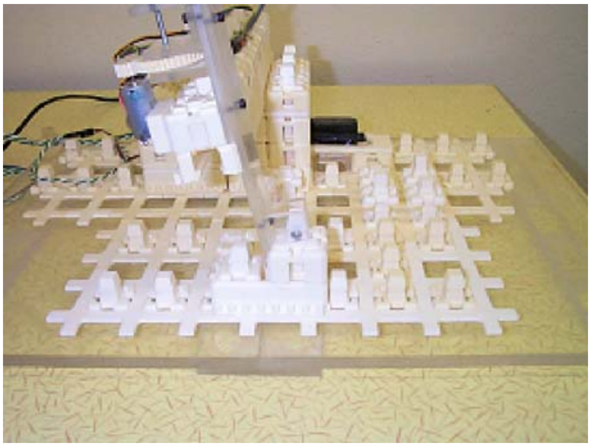
Deflection of the boom in the x-direction presented problems for sliding the far racks into place (Fig. 27). This would cause the sliding platform under the boom to jam. In addition to sideways deflection, the boom would also twist along the y-axis (Fig. 28). The constructor was able to place the near racks, but the far racks had to be placed by hand.



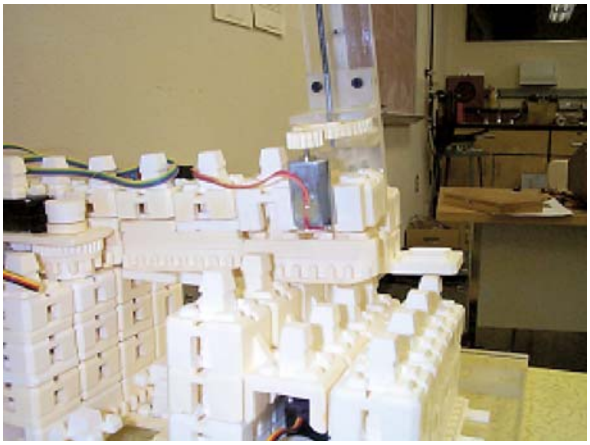
Self-replicating Robotic Systems, Figure 27
Sideways deflection of extended boom caused the sliding platform to jam

The torsion of the boom created problems in later assemblies steps as well. The twisting deflection of the boom would tend to misalign parts so that their snap-tangs would not engage with the parts below them. For this reason, the tool was mounted to the slide in a rotating bracket. This allowed the tool a small, passive amount of rotation about the y-axis. With this feature, the boom could still twist, but the parts would self-align and engage.

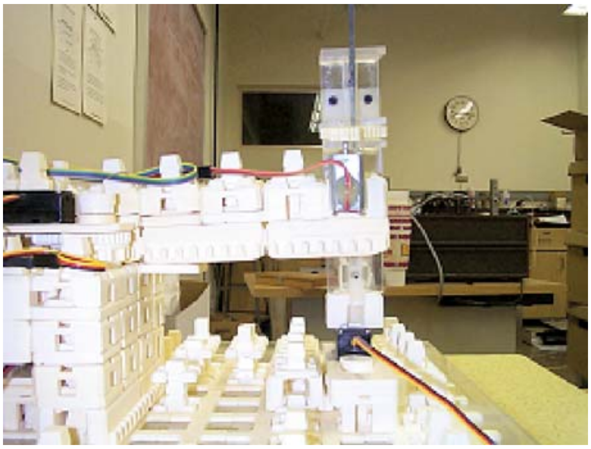
In many cases, the constructor could not develop the necessary force to assemble components. This was especially true when assembling components with multiple sets



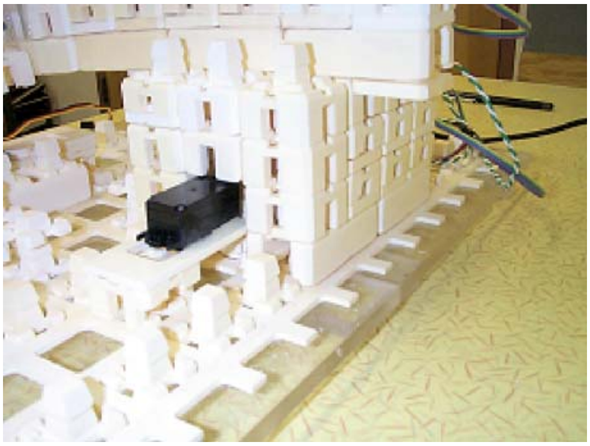
Self-replicating Robotic Systems, Figure 28
Torsion of boom



Self-replicating Robotic Systems, Figure 30
Separation of components within the boom



Self-replicating Robotic Systems, Figure 29
Upward deflection of boom



Self-replicating Robotic Systems, Figure 31
Separation of sliding platform and operating plane

of snap tangs, and in parts where the tangs were offset horizontally from the handle. Sometimes this occurred because the vertical linear actuator did not produce enough force, and sometimes it occurred because upward deflection of the boom caused the parts to misalign (Fig. 29). In a few cases, components in the boom would separate (Fig. 30). This only occurred when the boom was prevented from deflecting because it was engaged on the growing assembly.

The extended boom indirectly caused another problem. Recall that two of the handles in the tracks on the operating plane are modified as entry sites for racks. These sites do not help hold the platform to the plane. This, in combination with the high moment exerted on the platform by the extended boom, would cause the platform to separate from the plane and get stuck on adjacent, non-modified handles (Fig. 31).

Some aspects of the constructor worked well, such as the stage-by-stage assembly of the boom and the retrieval of parts from the storage site. The boom does not have to extend to reach the storage site, so deflections are not a problem and part retrieval was easy. The thickness of the tool tangs was found by trial and error – the tangs of a basic part were successively machined until the disassembly force fell within the correct range. Although no measurements were taken, the distribution of assembly and disassembly forces across various parts was fairly tight. The tool could lift all of the parts. Once parts were connected to an assembly, the tool could be separated from the connected parts in all cases.

In short, this initial concept of a desktop universal constructor provided us many valuable lessons about the importance of mechanical design, parts assembly and manufacturability issues that we are incorporating in future designs.

Discussion

Several self-replicating robot prototype have been constructed and tested. An autonomous prototype uses two light sensors in its navigation system to detect objects and also to track lines. Magnets and shape-constraining blocks are used to aid in aligning and interlocking the subsystems of the replica. As a result, the robot is capable of automatically assembling its replicas. All of the replicas are also capable of completing the same replicating process. The autonomous self-replicating system presented here has been recognized by authors of the book, "Kinematic Self-Replicating Machines" [25] that this prototype is the world's first fully functional autonomous self-replicating robot. Self-replicating circuits, as well as self-replicating robots whose controllers can be fully decomposed into basic parts have also been reviewed here.

Whereas von Neumann's architecture for self-replicating kinematic automata is the most widely known approach, it is not the only one. Self-reproduction by self-inspection in which a non-universal constructor "reads" an original device and "writes" a copy by executing a very small set of hardwired commands is an alternative. In our experience observing students attempting to build self-replicating devices, self-replication by self-inspection appears to be a more robust and less complicated alternative to the universal constructor.

Future Works

In previous sections we discussed a series of mechanical replicating prototypes in which programs are preloaded onto control computers which are then treated as one of several subsystems to be assembled. However, our ultimate goal is to develop a self-replicating robotic system capable of autonomously assembling its replicas from simple components using only electro-mechanical intelligence, i. e. a mechanical code, and transistor-based control circuits. This eliminates complicated electronic components, such as programmable micro-controllers, and makes the concept more appropriate for future space systems that can use in-situ resources for self-replication. The prototypes reviewed here are one step in this direction. Other related technologies are discussed in [33].

Acknowledgments

This work was made possible by support from the Royal Thai government, the Mellon Foundation, and the National Science Foundation under Grant ITS 0205466. The results and opinions expressed are solely those of the authors. We thank A. Cushing, Y. Kwon, K. Lee, M. Kutzer and all of the undergraduate students in the Mechatronics courses over the past five years who have designed several of the prototypes shown in this review. The work presented in Sect. "Towards a Universal Constructor" was performed at University of New Mexico with help from Professors G. Starr, H. Tran, and J. Wood.

Bibliography

1. Chirikjian GS, Pamecha A, Ebert-Uphoff I (1996) Evaluating Efficiency of Self-Reconfiguration in a Class of Modular Robots. J Robot Syst 13(5):317–338
2. Chirikjian GS, Suthakorn J (2002) Towards Self-Replicating Robots. In: Proceedings of the Eight International Symposium on Experimental Robotics (ISER), Italy, July 2002
3. Chirikjian GS, Zhou Y, Suthakorn J (2002) Self-Replicating Robots for Lunar Development. IEEE/ASME Trans Mechatron 7(4):462–472, December 2002
4. Eno S, Mace L, Liu J, Benson B, Raman K, Lee K, Moses M, Chirikjian GS (2007) Robotic Self-Replication in a Structured Environment without Computer Control. In: Proceedings of the 2007 IEEE CIRA. Piscataway
5. Freitas RA Jr (1980) A Self-Reproducing Interstellar Probe. J Br Interplanet Soc 33:251–264
6. Freitas RA Jr (1980) Report on the NASA/ASEE Summer Study on Advanced Automation for Space Missions. J Br Interplanet Soc 34:139–142
7. Freitas RA Jr (1983) Terraforming Mars and Venus Using Self-Replicating Systems. J Br Interplanet Soc 36:139–142
8. Freitas RA Jr, Merkle RC (2004) Kinematic Self-Replicating Machines. Landes Bioscience, Georgetown
9. Freitas RA Jr, Valdes F (1980) Comparison of Reproducing and Non-Reproducing Starprobe Strategies for Galactic Exploration. J Br Planet Soc 33:402–408
10. Freitas RA Jr, WP Gilbreath (eds) (1982) Advanced Automation for Space Missions. In: Proceedings of the 1980 NASA/ASEE summer study. Replicating Systems Concepts: Self-Replicating Lunar Factory and Demonstration, NASA, Scientific and Technical Information Branch (Conference Publication 2255), US Government Printing Office, Washington DC, chap 5
11. Hastings WA, Labarre M, Viswanathan A, Lee S, Sparks D, Tran T, Nolin J, Curry R, David M, Huang S, Shuthakorn J, Zhou Y, Chirikjian GS (2004) A minimalist parts manipulation system for a self replicating electromechanical circuit. IMG'04, Genoa, Italy, July 1–2, 2004
12. Hosokawa K, Fujii T, Kaetsu H, Asama H, Kuroda HY, Endo I (1999) Self-organizing collective robots with morphogenesis in a vertical plane. JSME Int J Ser C-Mechanical Syst Mach Elements Manuf 42(1):195–202
13. Jacobson H (1958) On Models of Reproduction. Am Scient 46:255–284

14. Kotay K, Rus D, Vona M, McGray C (1998) The Self-reconfiguring Molecule: Design and Control Algorithms. WAFR'98: Proceedings of the third workshop on the algorithmic foundations of robotics. Publisher A.K. Peters, Ltd. Wellesley
15. Lee K, Chirikjian GS (2007) Robotic Self-Replication. Low Complexity Parts. IEEE Robotics and Automation Magazine 14(4):34–43. Published by IEEE, Piscataway
16. Lee K, Moses M, Chirikjian GS (2008) Robotic Self-Replication in Structured Environments: Physical Demonstrations and Complexity Measures. Int J Robot Res 27(3-4):387–401. doi:10.1177/0278364907084982
17. Lipson H, Pollack B (2000) Automatic design and Manufacture of Robotic Lifeforms. Nature 406:974–978
18. Liu A, Sterling M, Kim D, Pierpont A, Schlothauer A, Moses M, Lee K, Chirikjian GS (2007) A Memoryless Robot that Assembles Seven Subsystems to Copy Itself. In: Proceedings of the 2007 IEEE ISAM, Piscataway
19. Moore EF (1956) Artificial Living Plants. Scient Am 195:118–126
20. Moses M (2001) A Physical Prototype of a Self-Replicating Universal Constructor. M.S. University of New Mexico
21. Murata S, Kurokawa H, Kokaji S (1994) Self-Assembling Machine. In: Proceedings of the 1994 IEEE International Conference on Robotics and Automation. Piscataway, San Diego, CA, pp 441–448
22. Penrose LS (1959) Self-Reproducing Machines. Sci Am 200(6):105–114
23. Pesavento U (1995) An implementation of von Neumann's self-reproducing machine. Artif Life J 2(4):337–354
24. Pfeifer R, Kunz H, Weber MM, Thomas D (2001) Lecture of the Artificial Life. http://www.ifi.unizh.ch/ailab/teaching/AL01/chap7.pdf. Dept of Information Technology, University of Zurich. Accessed 19 June 2008
25. Rebek J Jr (1994) Synthetic Self-Replicating Molecules. Sci Am 271(1):48–55
26. Sipper M (1998) Fifty Years of Research on Self-Replication: An Overview. Artif Life 4(3):237–257
27. Suthakorn J (2004) Paradigm for Service Robotics. Ph D Dissertation. Johns Hopkins University
28. Suthakorn J, Kwon YT, Chirikjian GS (2003) A Semi-Autonomous Replicating Robotic System. In: Proceedings of the 2003 IEEE/ASME International Symposium on Computational Intelligence for Robotics and Automation (CIRA), Kobe, Japan, Piscataway
29. Suthakorn J, Zhou Y, Chirikjian GS (2002) Self-Replicating Robots for Space Utilization. In: Proceedings of the 2002 Robophere workshop on Self Sustaining Robotic Ecologies, NASA Ames Research Center, California, NASA Ames Research Center
30. Tiesenhausen GV, Darbro WA (1980) Self-Replicating Systems – A Systems Engineering Approach. In: Technical Memorandum: NASA TM-78304, Washington DC
31. Von Neumann J, Burks AW (1966) Theory of Self-Reproducing Automata. University of Illinois Press,Champaign
32. Whitesides GM (1995) Self-Assembling Materials. Sci Am 273(3):146–149
33. Yim M, Shen WM, Salemi B, Rus D, Moll M, Lipson H, Klavins E, Chirikjian GS (2007) Modular self-reconfigurable robot systems – Challenges and opportunities for the future. IEEE Robotics Autom Mag 14(1):43–52
34. Yim M, Zhang Y, Lamping J, Mao E (2001) Distributed Control for 3D Metamorphosis. Auton Robots 10:41–56

Self-Replication and Cellular Automata

GIANLUCA TEMPESTI[1], DANIEL MANGE[2], ANDRÉ STAUFFER[2]

[1] University of York, York, UK

[2] Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Article Outline

Glossary

Cellular automaton A cellular automaton (CA) is a mathematical framework modeling an array of *cells* that interact locally with their neighbors. In this *cellular space*, each cell has a set of *neighbors*, cells have *values* or *states*, all the cells update their values simultaneously at discrete *time steps* or *iterations*, and the new state of a cell is determined by the current state of its neighbors (including itself) according to a local *function* or *rule*, identical for all cells. In the article, the term is extended to account for systems that introduce variations to the basic definition (for example, systems where cells do not update simultaneously or do not have the same set of rules in every cell).

Following the historical pattern, in the article the same term is also used to refer to an object or structure built within the cellular space, i. e., a set of cells in a particular, usually active, state (overlapping with the definition of *Configuration*).

Configuration A set of cells in a given state at a given time. Usually, but not always, the term refers to the state of all the cells in the entire space. The *initial configuration* is the state of the cells at time $t = 0$.

Self-replication The process whereby a cellular automaton configuration creates a copy of itself in the cellular space. Incidentally, you will note that in the article we use the terms self-replication and self-reproduction interchangeably. In reality, the two terms are not really synonyms: self-reproduction is more properly applied to the reproduction of organisms, while self-replication concerns the cellular level. The more correct term to use in most cases would probably be self-replication, but since von Neumann favored self-reproduction, we will ignore the distinction.

Self-reproduction See *Self-Replication*

Construction The process that occurs when one or more cells, initially in the inactive or *quiescent* state are assigned an active state (in the context of this article, by the self-replicating structure).

Definition of the Subject

Machine self-replication, besides inspiring numerous fictional books and movies, has long been considered a powerful paradigm to allow artifacts, for example, to survive in hostile environments (such as other planets) or to operate more efficiently by creating *populations* of machines working together to achieve a given task. Where the self-replication of *computing* machines is concerned, other motivations can also come into play, related to concepts such as fault tolerance and self-organization.

Cellular automata have traditionally been the framework of choice for the study of self-replicating computing machines, ever since they were used by John von Neumann, who pioneered the field in the 1950s. In this context, self-replication is seen as the process whereby a configuration in the cellular space is capable of creating a copy of itself in a different location.

As a mathematical framework, CA allow researchers to study the mechanisms required to achieve self-replication in a simplified environment, in view of eventually applying this process to real-world systems, either to electronics or, more generally, to computing systems.

Introduction

The self-replication of computing systems is an idea that dates back to the very origins of electronics. One of the pioneers of the field, John von Neumann, was among the first to investigate the possibility of creating machines capable of self-replication [1] with the purpose of achieving reliability through the redundant operation of "populations" of computing machines.

Throughout the more than 50 years since von Neumann's seminal work, research on this topic has gone through several transformations. While interest in applying self-replication to electronic systems waned because of technological hurdles, the field of Artificial Life, starting with the pioneering work of Chris Langton [14], began studying this process in the more general context of achieving life-like properties in artificial systems.

Throughout its long history, cellular automata (CA) have remained one of the environments of choice to study how self-replication can be applied to computing systems. In general, researchers in the domain (including von Neumann) have never regarded CA as the environment in which self-replication would be ultimately applied. Rather, CA have traditionally provided a useful platform to test the complexity of self-replication at an early stage, in view of eventually applying this process to real-world systems, either to electronics or, more generally, to computing systems.

Of course, the concept of self-replication has been applied to artificial systems in contexts other than computing. A classic example is the 1980 NASA study by Robert Freitas Jr. and Ralph Merkle [10] (recently expanded in a remarkable book [11]), where self-replication is used as a paradigm for efficiently exploring other planets. However, this kind of self-replication, applied to physical machines rather then computing systems, does not commonly make use of cellular automata and is beyond the scope of this article.

Following the historical progress of self-replication in cellular automata (derived in part from [40]), we will first examine in some detail von Neumann's seminal work (Sect. "Von Neumann's Universal Constructor"). Then, the use of self-replication as an Artificial Life paradigm will be discussed (Sect. "Self-Replication for Artificial Life") before dealing with some of the latest advances in the field in Sect. "Other Approaches to Self-Replication".

Von Neumann's Universal Constructor

Many of the existing approaches to the self-replication of computing systems are essentially derived from the work of John von Neumann [1], who pioneered this field of research in the 1950s.

Von Neumann, confronted with the lack of reliability of computing systems, turned to nature to find inspiration in the design of fault-tolerant computing machines. Let us remember that the computers von Neumann was familiar with were based on vacuum-tube technology, and that vacuum tubes were much more prone to failure than modern transistors. Moreover, since the writing and the execution of complex programs on such systems represented many hours (if not many days) of work, the failure of a system had important consequences in wasted time and effort.

In particular, Von Neumann investigated self-replication as a way to design and implement digital logic devices. Unfortunately, the state of the art in the fifties restricted von Neumann's investigations to a purely theoretical level, and the work of his successors mirrored this constraint.

Indeed, it is not until fairly recently that some of the technological problems associated with the implementation of such a process in silicon have been resolved with the introduction of new kinds of electronic devices (see Sect. "Other Approaches to Self-Replication").

In this section, we will analyze von Neumann's research on the subject of self-replicating computing machines, and in particular his universal constructor, a self-replicating cellular automaton [44].

Von Neumann's Self-Replicating Machines

Natural systems are among the most reliable complex systems known to man, and their reliability is a consequence not of any particular robustness of the individual cells (or organisms), but rather of their extreme redundancy. The basic natural mechanism which provides such reliability is *self-reproduction*, both at the cellular level (where the survival of a single organism is concerned) and at the organism level (where the survival of the species is concerned).

Thus von Neumann, drawing inspiration from natural systems, attempted to develop an approach to the realization of self-replicating computing machines (which he called *artificial automata*, as opposed to natural automata, that is, biological organisms). In order to achieve his goal, he imagined a series of five distinct models for self-reproduction ([44], pp. 91–99):

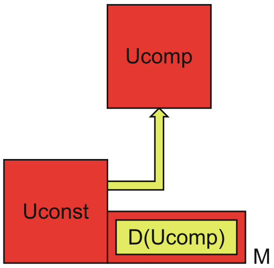
1. The *kinematic* model, introduced by von Neumann on the occasion of a series of five lectures given at the University of Illinois in December 1949, is the most general. It involves structural elements such as sensors, muscle-like components, joining and cutting tools, along with logic (switch) and memory elements. Concerning, as it does, physical as well as electronic components, its goal was to define the bases of self-replication, but was not designed to be implemented.
2. In order to find an approach to self-replication more amenable to a rigorous mathematical treatment, von Neumann, following the suggestion of the mathematician S. Ulam, developed a *cellular* model. This model, based on the use of cellular automata as a framework for study, was probably the closest to an actual realization. Even if it was never completed, it was further refined by von Neumann's successors and was the basis for most further research on self-replication.
3. The *excitation-threshold-fatigue* model was based on the cellular model, but each cell of the automaton was replaced by a neuron-like element. Von Neumann never defined the details of the neuron, but through a careful analysis of his work, we can deduce that it would have borne a fairly close relationship to today's simplest artificial neural networks, with the addition of some features which would have both increased the resemblance to biological neurons and introduced the possibility of self-replication.
4. For the *continuous* model, von Neumann planned to use differential equations to describe the process of self-reproduction. Again, we are not aware of the details of this model, but we can assume that von Neumann planned to define systems of differential equations to describe the excitation, threshold and fatigue properties of a neuron. At the implementation level, this would probably correspond to a transition from purely digital to analog circuits.
5. The *probabilistic* model is the least well-defined of all the models. We know that von Neumann intended to introduce a kind of automaton where the transitions between states were probabilistic rather than deterministic. Such an approach would allow the introduction of mechanisms such as mutation and thus of the phenomenon of evolution in artificial automata. Once again, we cannot be sure of how von Neumann would have realized such systems, but we can assume they would have exploited some of the same tools used today by genetic algorithms.

Of all these models, the only one von Neumann developed in some detail was the cellular model. Since it was the basis for the work of his successors, it deserves to be examined more closely.

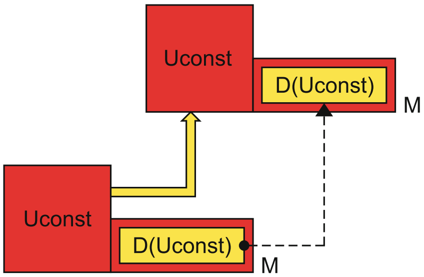
Von Neumann's Cellular Model

In von Neumann's work, self-reproduction is always presented as a special case of universal construction, that is, the capability of building any machine given its description (Fig. 1). This approach was maintained in the design of his cellular automaton, which is therefore much more than a self-replicating machine. The complexity of its purpose is reflected in the complexity of its structure, based on three separate components:

1. A *memory tape*, containing the description of the machine to be built, in the form of a one-dimensional string of elements. In the special case of self-reproduction, the memory contains a description of the universal constructor itself (Fig. 2). It is interesting to note that the memory of von Neumann's automaton bears a strong resemblance to the biological genome. This resemblance is even more remarkable when considering that the structure of the genome was not discovered until after the death of von Neumann.



Self-Replication and Cellular Automata, Figure 1
Von Neumann's universal constructor `Uconst` can build a specimen of any machine (e. g., a universal Turing machine `Ucomp`) given its description `D(Ucomp)`



Self-Replication and Cellular Automata, Figure 2
Given its own description `D(Uconst)`, von Neumann's universal constructor is capable of self-replication

2. The *constructor* itself, a very complex machine capable of reading the memory tape and interpreting its contents.
3. A *constructing arm*, directed by the constructor, used to build the offspring (i. e., the machine described in the memory tape). The arm moves across the space and sets the state of the elements of the offspring to the appropriate value.

The implementation as a cellular automaton is no less complex. Each element has 29 possible states, and thus, since the next state of an element depends on its current state and that of its four cardinal neighbors, $29^5 = 20,511,149$ transition rules are required to exhaustively define its behavior.

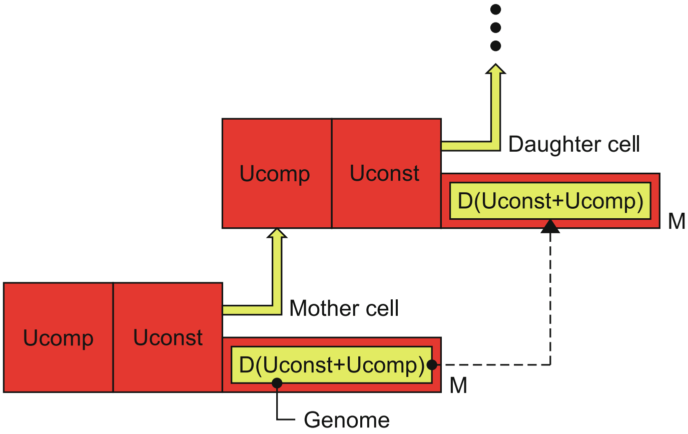
If we consider that the size of von Neumann's constructor is of the order of several hundred thousand elements, we can easily understand why a hardware realization of such a machine is not really feasible. In fact, as part of our research, we did realize a hardware implementation of a set of elements of von Neumann's automaton [3,35]. By carefully designing the hardware structure of each element, we were able to considerably reduce the amount of memory required to host the transition rules. Nevertheless, our system remains a demonstration unit, as it consists of a few elements only, barely enough to illustrate the behavior of a tiny subset of the entire machine.

It is also worth mentioning that von Neumann went one step further in the design of his universal constructor. If we consider the universal constructor from a biological viewpoint, we can associate the memory tape with the genome, and thus the entire constructor with a single cell (which would imply a parallel between the automaton's elements and molecules). However, the constructor, as we have described it so far, has no functionality outside of self-reproduction. Von Neumann recognized that a self-replicating machine would require some sort of functionality to be interesting from an engineering point of view, and postulated the presence of a *universal computer* (in practice, a universal Turing machine, an automaton capable of performing any computation) alongside the universal constructor (Fig. 3). Von Neumann's constructor can thus be regarded as a unicellular organism, containing a genome stored in the form of a memory tape, read and interpreted by the universal constructor (the mother cell) both to determine its operation and to direct the construction of a complete copy of itself (the daughter cell).

Von Neumann's Successors

The extreme size of von Neumann's universal constructor has so far prevented any kind of physical implementation (apart from the small demonstration unit we mentioned). But further, even the simulation of a cellular automaton of such complexity was far beyond the capability of early computer systems. Today, such a simulation is starting to be conceivable. Umberto Pesavento, a young Italian high school student, developed a simulation of von Neumann's entire universal constructor [27]. The computing power available did not allow him to simulate either the entire self-replication process (the length of the memory tape needed to describe the automaton would have required too large an array) or the Turing machine necessary to implement the universal computer, but he was able to demonstrate the full functionality of the constructor.

Considering the rapid advances in computing power of modern computer systems, we can assume that a complete simulation could be envisaged with today's technology. In fact, an effort is currently under way [4] to implement a complete specimen of the constructor. To achieve this goal, Buckley is revisiting and analyzing in detail the operation of the constructor. To give an idea of the scope of this work, Buckley's results indicate that the interpretercopier (without the tape) is bounded by a region of 751 × 1048 = 787.048 cells.



Self-Replication and Cellular Automata, Figure 3
By extension, von Neumann's universal constructor can include a universal computer and still be capable of self-replication

The impossibility of achieving a physical realization did not however deter some researchers from trying to continue and improve von Neumann's work [2,15,22]. Arthur Burks, for example, in addition to editing von Neumann's work on self-replication [5,44], also made several corrections and advances in the implementation of the cellular model. Codd [9], by altering the states and the transition rules, managed to simplify the constructor by a considerable degree. Vitanyi [43] studied the possibility of introducing sexual reproduction in von Neumann's approach. However, without in any way lessening these contributions, we can say that no major theoretical advance in the research on self-reproducing automata occurred until C. Langton, in 1984, opened a second stage in this field of research.

Self-Replication for Artificial Life

While the *implementation* of von Neumann's universal constructor faced insurmountable (at the time) technological hurdles, the same could not be said of the *theoretical* contribution that his approach represented as an attempt to study a biologically-inspired process within the world of computing systems.

In this context, the main drawback of von Neumann's work lay in the inability to achieve self-replication without resorting to an extremely complex simulation of a complete machine. Von Neumann's Universal Constructor was so complex because it tried to implement self-reproduction as a particular case of construction universality, i. e. the capability of constructing any other automaton, given its description. C. Langton approached the problem somewhat differently, by attempting to define the simplest cellular automaton, commonly known as *Langton's loop* [14], capable exclusively of self-reproduction.

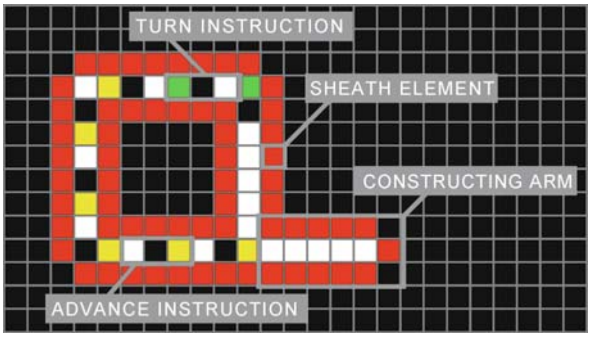
Langton's loop had a major impact on research in self-replication by introducing a new way to think about this process in more "abstract" terms as a study of the application of biologically-inspired mechanisms to computing, exemplifying the field known as *Artificial Life*. In this context, rather than the replicating machine, it is the process of self-replication itself that becomes the object of study.

This novel approach generated research that can be considered fundamentally different from that of von Neumann and started discussion on topics such as the analogy with cellular division and with the reproduction of individuals in a population (e.g, in [17], Section V.B), the difference between trivial and non-trivial self-replication in cellular automata (e. g., in automata such as those described in [16]), or the connections between evolution and self-replication (e.g, in the work of Sayama [17]).

Langton's Loop

As a consequence of his approach, Langton's Loop is orders of magnitude simpler than von Neumann's constructor. In fact, it is loosely based on one of the simplest organs (an organ in this context can be seen as a self-supporting structure capable of a single sub-task) in Codd's automaton: the *periodic emitter* (itself derived from von Neumann's periodic pulser), a relatively simple structure capable of generating a repeating string of a given sequence of pulses.

Langton's loop (Fig. 4) is named after the dynamic storage of data inside a square sheath (red in the figure). The data is stored as a sequence of instructions for directing the constructing arm, coded in the form of a set of three states. The data turns counterclockwise in permanence within the sheath, creating a loop.



Self-Replication and Cellular Automata, Figure 4
The initial configuration of Langton's Loop (iteration 0)

The two instructions in Langton's loop are extremely simple. One instruction (uniquely identified by the yellow element in the figure) tells the arm to advance by one position (Fig. 5), while the other (green in the figure) directs the arm to turn 90 degrees to the left (Fig. 6). Obviously, after three such turns, the arm has looped back on itself (Fig. 7), at which stage a messenger (the pink element) starts the process of severing the connection between the parent and the offspring, thus concluding the replication process.

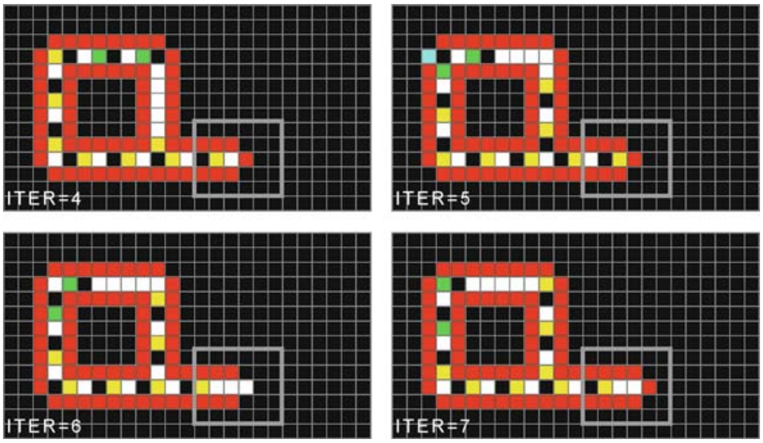
Once the copy is over, the parent loop proceeds to construct a second copy of itself in a different direction (to the north in this example), while the offspring itself starts to reproduce (to the east in this example). The sequential nature of the self-reproduction process generates a spiraling pattern in the propagation of the loop through space (Fig. 8): as each loop tries to reproduce in the four cardinal directions, it finds the place already occupied either by its parent or by the offspring of another loop, in which case it dies (the data within the loop is destroyed).

Langton's loop uses 8 states for each of the 86 non-quiescent cells making up its initial configuration, a 5-cell neighborhood, and a few hundred transition rules (the exact number depends on whether default rules are used and whether symmetric rules are included in the count). Further simplifications to Langton's automaton were introduced by Byl [6], who eliminated the internal sheath and reduced the number of states per cell, the number of transition rules, and the number of non-quiescent cells in the initial configuration. Reggia et al. [29] managed to remove also the external sheath, thus designing the smallest self-replicating loop known to date. Given their modest complexity, at least relative to von Neumann's automaton, all of the mentioned automata have been thoroughly simulated.

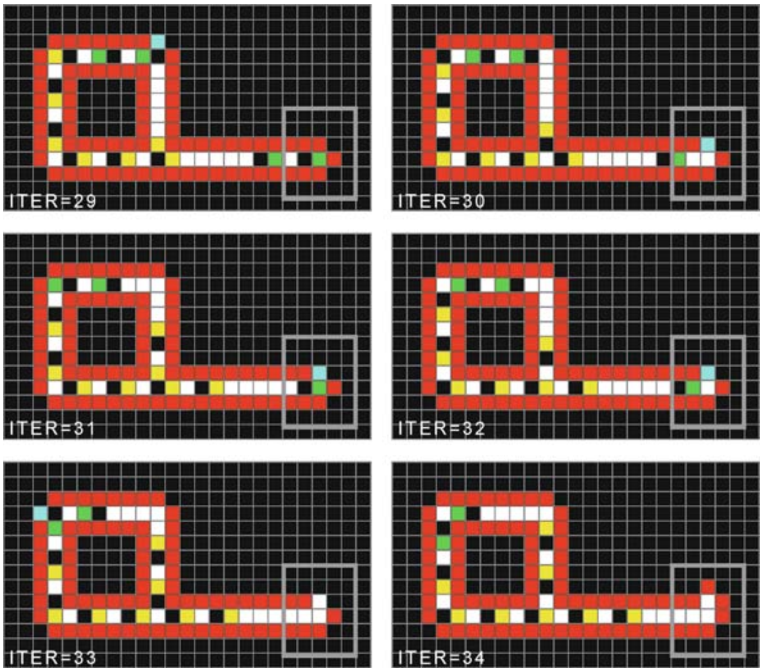
Langton's loop has been used as the basis for several approaches, mostly aimed at studying the properties of self-replication within a cellular system in the context of artificial life. Sayama [32] introduced structural dissolution (whereby a loop can destroy itself, in addition to replicating) to obtain colonies of loops that are dynamically stable and exhibit a potentially evolvable behavior. Nehaniv [21] extended Langton's approach to asynchronous cellular automata, while Sipper [34] developed a self-replicating loop in a non-uniform CA (i. e., a CA where the transition rules are not necessarily identical in all cells).

Perrier's Loop

In the context of applying self-reproduction to the replication of computing machines, and hence return to von Neumann's original goals, the main weakness of Langton's loop resides in the absence of any functionality beyond self-reproduction itself. To overcome this limitation, Per-



Self-Replication and Cellular Automata, Figure 5
The constructing arm advances by one space



Self-Replication and Cellular Automata, Figure 6
The constructing arm turns 90 degrees to the left

rier and Zahnd developed a relatively complex automaton (Fig. 9) in which a two-tape Turing machine was appended to Langton's loop [26].

This automaton exploits Langton's loop as a sort of "carrier" (Fig. 10): the first operation of Perrier's loop is to allow Langton's loop to build a copy of itself (iteration 128: note that the copy is limited to one dimension, since the second dimension is taken up by the Turing machine). The main function of the offspring is to determine the location of the copy of the Turing machine (iteration 134). Once the new loop is ready, a "messenger" runs back to the parent loop and starts to duplicate the Turing machine (iterations 158 and 194), a process completely disjoint from the operation of the loop. When the copy is finished (iteration 254), the same messenger activates the Turing machine in the parent loop (the machine had to be inert during the replication process in order to obtain a perfect copy). The process is then repeated in each offspring until the space is filled (iteration 720: as the automaton exploits Langton's loop for replication, meeting the boundary of the array causes the last machine to crash).

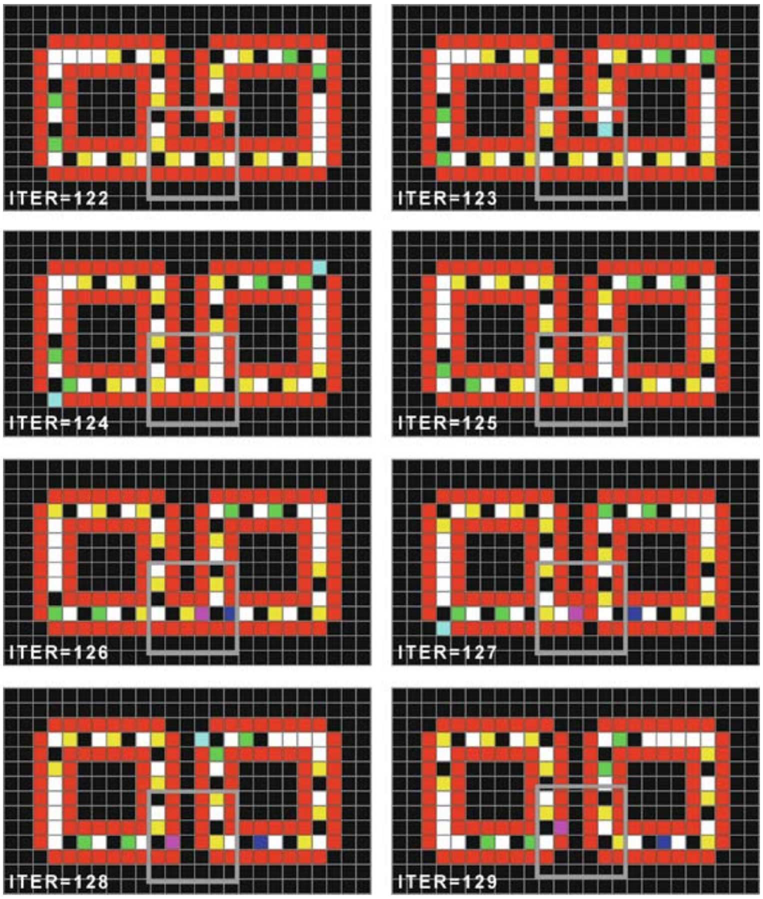
Perrier's automaton implements a self-replicating Turing machine, a powerful construct which is unfortunately handicapped by its complexity: in order to implement a Turing machine, the automaton requires a very considerable number of additional states (more than 60), as well as an important number of additional transition rules. This kind of complexity, while still relatively minor compared to von Neumann's universal constructor, is nevertheless too important to be really considered for an actual implementation.

Tempesti's Loop

Always in the context of achieving self-reproduction of computing machines, and beside the lack of functionality mentioned in Subsect. "Perrier's Loop", another problem of Langton's loop is that it is not well adapted to finite CA arrays. Its self-reproduction mechanism assumes that there is enough space for a copy of the loop, and the entire loop becomes nonfunctional otherwise (Fig. 8).

To overcome this limitation and move a step closer to the realization of self-replicating machines, we developed a self-replicating loop designed specifically to exist in a finite, but arbitrarily large, space, and at the same time capable, unlike Langton's loop, to have a functionality in addition to self-replication.

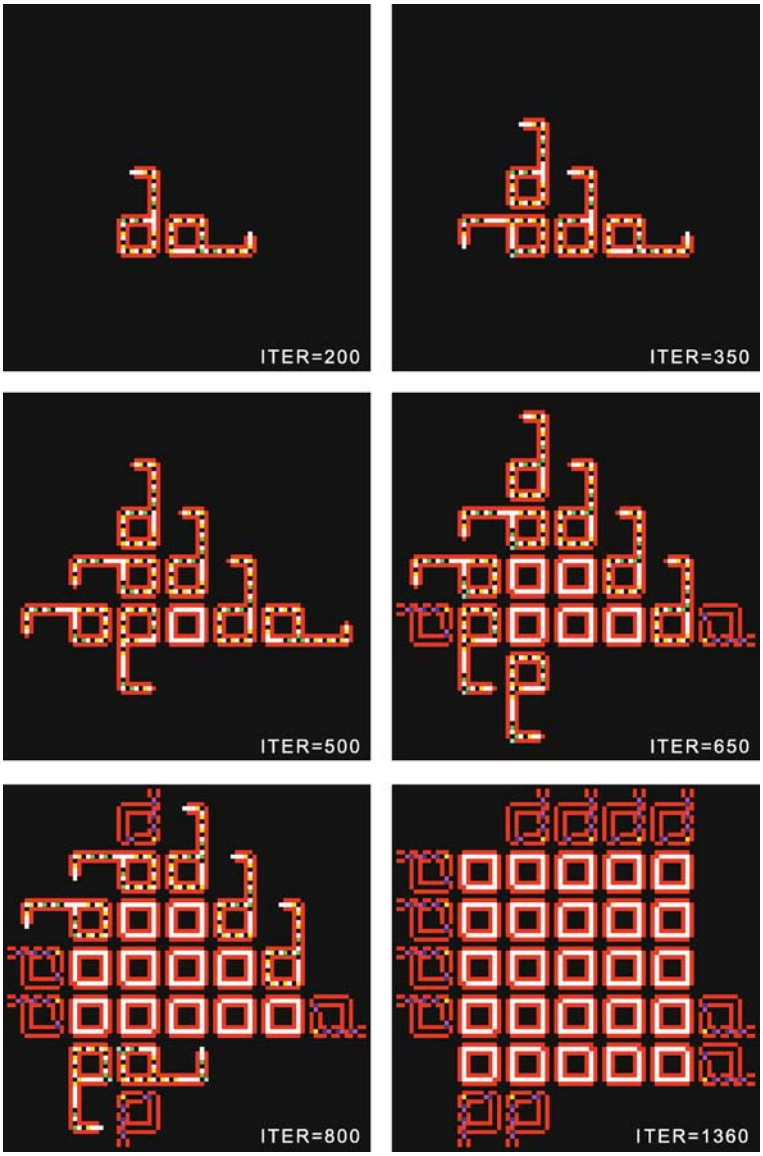
In designing our self-replicating automaton [39,40], we did maintain some of the more interesting features of Langton's loop. In particular, we preserved the structure based on a square loop to dynamically store information. Such storage is convenient in CA because of the locality of



Self-Replication and Cellular Automata, Figure 7
The copy is complete and the connection from parent to offspring is severed

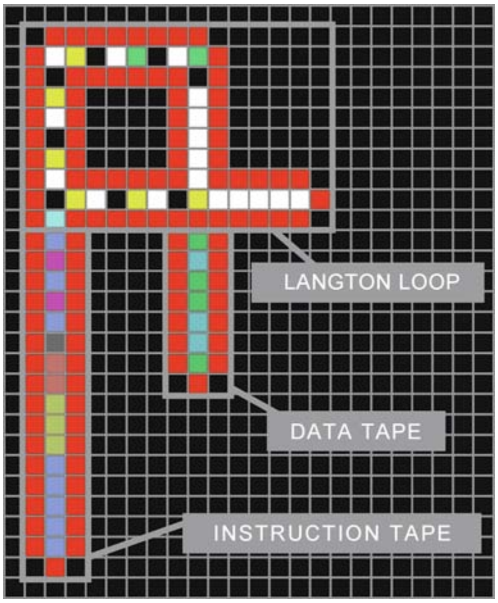
the rules. Also, we maintained the concept of constructing arm, in the tradition of von Neumann and his successors, even if we introduced considerable modifications to its structure and operation. While preserving some of the more interesting features of Langton's loop, we nevertheless introduced some basic structural alterations (Fig. 11):

- As in Byl's version of Langton's loop, we use only one sheath. In addition, four *gate elements* (in the same state as the sheath) at the four corners of the automaton enable or disable the replication process.
- We extend four constructing arms in the four cardinal directions at the same time, and thus create four copies of the original automaton in the four directions in parallel. When the arm meets an obstacle (either the boundary of the array or an existing copy of the loop), it simply retracts and puts the corresponding gate element in the closed position.
- The arm does not immediately construct the entire loop. Rather, it constructs a sheath of the same size as the original. Once the sheath is ready, the data circulating in the loop is duplicated and the copy is sent along the constructing arm to wrap around the new sheath. When the new loop is completed, the constructing arm retracts and closes the gate.
- As a consequence, we use only four of the elements circulating in the loop to control the self-replication process. The others can be used as a "program", i. e., a set of states with their own transition rules which will then be applied alongside the self-reproduction to execute some function.
- Unlike Langton's loop, our loop does not "die" once duplication is achieved, as the circulating data remains untouched by the self-reproduction process. This feature is a requirement for implementing functions which work after the copy has finished.



Self-Replication and Cellular Automata, Figure 8
Propagation pattern of Langton's loop

We use a 9-element neighborhood (the element itself plus its 8 neighbors) and the basic configuration of the loop (Fig. 11) requires five states plus at least one data state. State 0 (black) is the *quiescent* state: it represents the inactive background. State 1 (white) is the *sheath* state, that is the state of the elements making up the sheath and the four gates. State 2 (red) is the *activation* state or *control* state. The four gate elements are in state 2, as are the messengers which will be used to command the constructing arm and the tip of the constructing arm itself for the first phase of construction, after which the tip of the arm will switch to state 3 (light blue), the *construction* state. State 3 will construct the sheath that will host the offspring, signal the parent loop that the sheath is ready, and lead the duplicated data to the new loop. State 4 (green), the destruction state, will destroy the constructing arm once the copy is completed. In addition to these states, two additional *data* states (light and dark grey) represent the information stored in the loop. In this example, they are inactive, while the next section describes a loop where they are used to store an executable program.



Self-Replication and Cellular Automata, Figure 9
A two-tape Turing machine appended to Langton's loop (iteration 0)

The initial configuration is in the form of a square loop wrapped around a sheath. The size of the loop is variable, and for our example is set to 8 × 8. Once the iterations begin, the data starts turning counterclockwise around the loop. Nothing happens until the first control element (red) reaches a corner of the loop, where it checks the status of the gate. Since the gate is open, the control element splits into two identical elements: the first continues turning around the loop, while the second starts extending the arm (Fig. 12).

The arm advances automatically by one position every two iterations. Once the arm has started extending, each control element that arrives to a corner will again split and one of the copies will start running along the arm, advancing twice as fast. When the first of these messengers reaches the tip of the arm, the tip, which was until then in state 2, passes to state 3 and continues to advance at the same speed. This transformation tells the arm that it has reached the location of the offspring loop and to start constructing the new sheath. The next three messengers will force the tip of the arm to turn left, while the fourth causes the sheath to close upon itself (Fig. 13). It then runs back along the arm to signal to the original loop that the new sheath is ready. Once the return signal arrives at the corner of the original loop, it causes a copy of the data in the loop to run along the arm and wrap itself around the new sheath. Once the second copy has completed the loop (Fig. 14), it sends a destruction signal (green) back along the arm. The signal will destroy the arm until it reaches the corner of the original loop, where it closes the gate to avoid further copies.

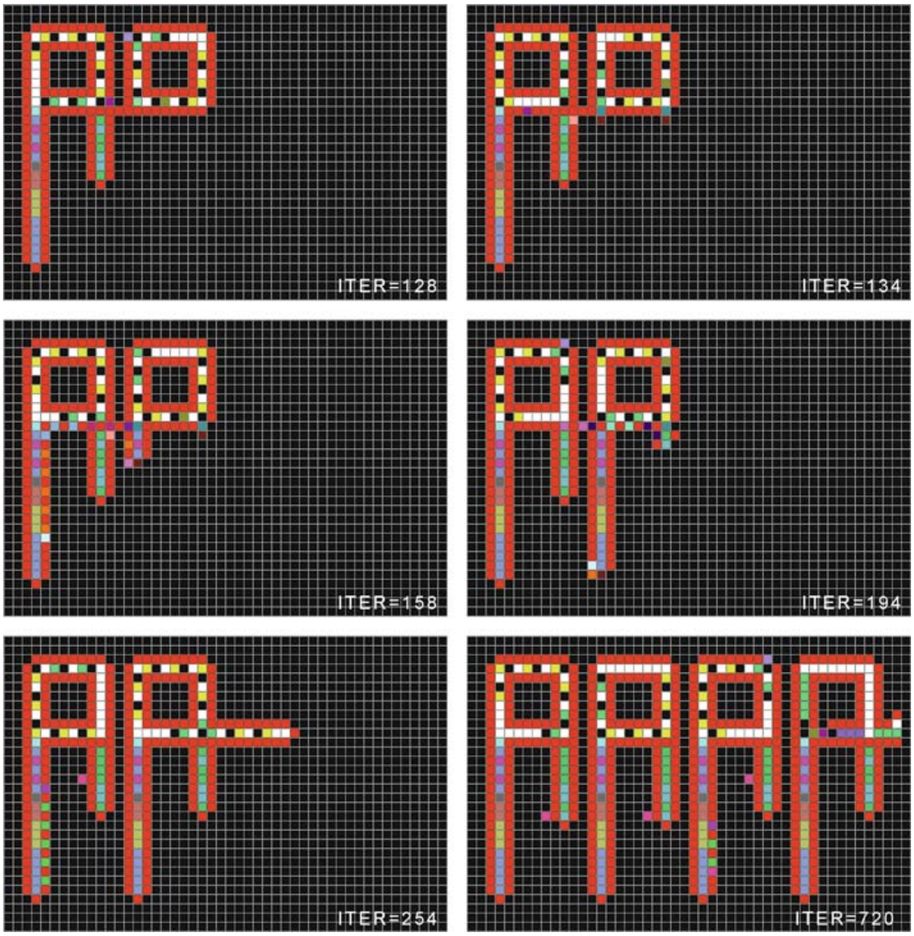
After 121 time periods the gates of the original automaton will be closed and it will enter an inactive state, with the understanding that it will be ready to reproduce itself again should the gates be opened. The main advantage of the new mechanism is that it becomes relatively simple to retract the arm if an obstacle (either the boundary of the array or another loop) is encountered, and therefore our loop is perfectly capable of operating in a finite space.

In Fig. 15, we illustrate an example of how the data states can be used to carry out operations alongside self-reproduction. The operation in question is the construction of three letters, LSL (the acronym of Logic Systems Laboratory, where the research was made), in the empty space inside the loop. Obviously this is not a very useful operation from a computational point of view, but it is a far from trivial construction task which should suffice to demonstrate the capabilities of the automaton.

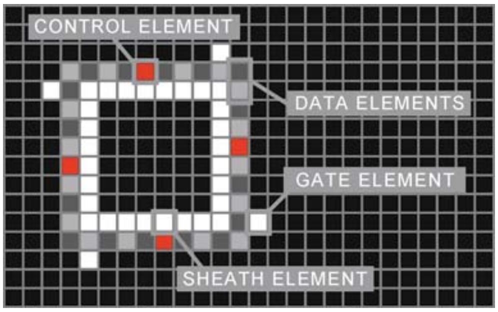
As should be obvious, while our loop owes to von Neumann the concept of constructing arm and to Langton (and/or Codd) the basic loop structure, it is in fact a very different automaton, endowed with some of the properties of both. We have seen that von Neumann's automaton is extremely complex, while Langton's loop is very simple. The complexity of our automaton is more difficult to estimate, as it depends on the data circulating in the loop. The number of non-quiescent elements making up the initial configuration depends directly on the size of the circulating program. The more complex (i. e. the longer) the program, the larger the automaton (it should be noted, however, that the complexity of the self-reproduction process does not depend on the size of the loop). The number of transition rules is obviously a function of the number of data states: in the basic configuration (i. e., one data state), the automaton needs 692 rules (173 rules rotated in the four directions), assuming that, by default, all elements remain in the same state. The complexity of the basic configuration is therefore in the same order as that of Langton's and Byl's loops, with the proviso that it is likely to increase drastically if the data in the loop is used to implement a complex function.

Other Approaches to Self-Replication

Von Neumann's and Langton's structures represent the main landmarks in the study of self-replication in com-



Self-Replication and Cellular Automata, Figure 10
Self-replication of the Turing machine

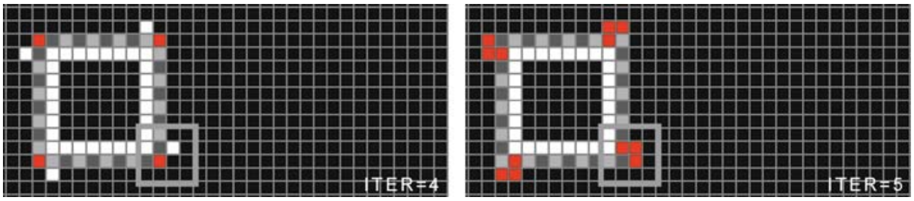


Self-Replication and Cellular Automata, Figure 11
The initial configuration of our loop (iteration 0)

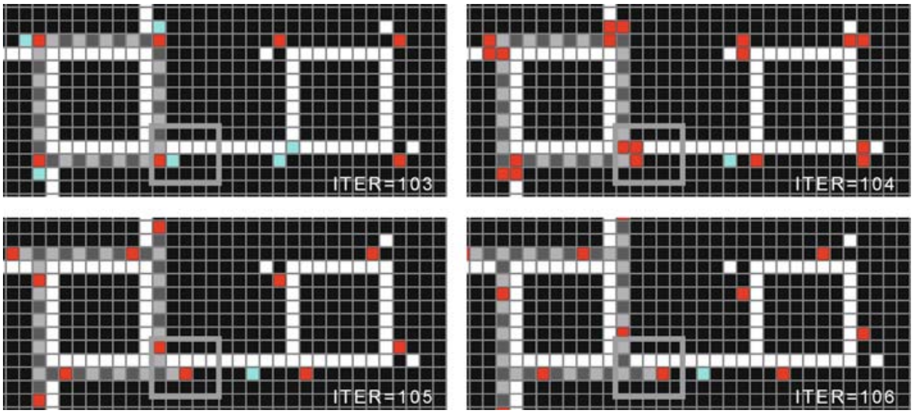
puting machines. It can safely be said that all other approaches refer, directly or indirectly, to these two systems. However, there exist some approaches to self-replication that cannot be easily reduced to simple variations on one of these two themes, either because they specifically take into consideration some issues that are not addressed by Langton and von Neumann, or because they occur in environments that are considerably different from the two original approaches.

In this section, we will deal in depth with one example, the *Tom Thumb algorithm* that, while referring back to von Neumann insofar as its goal is the implementation of self-replicating logic circuits, is specifically designed to operate efficiently in the kind of digital devices that are available today. The algorithm approaches cellular automata from a slightly unconventional angle [37], with the objective of a hardware realization of self-replication within a programmable logic device, or FPGA [41].

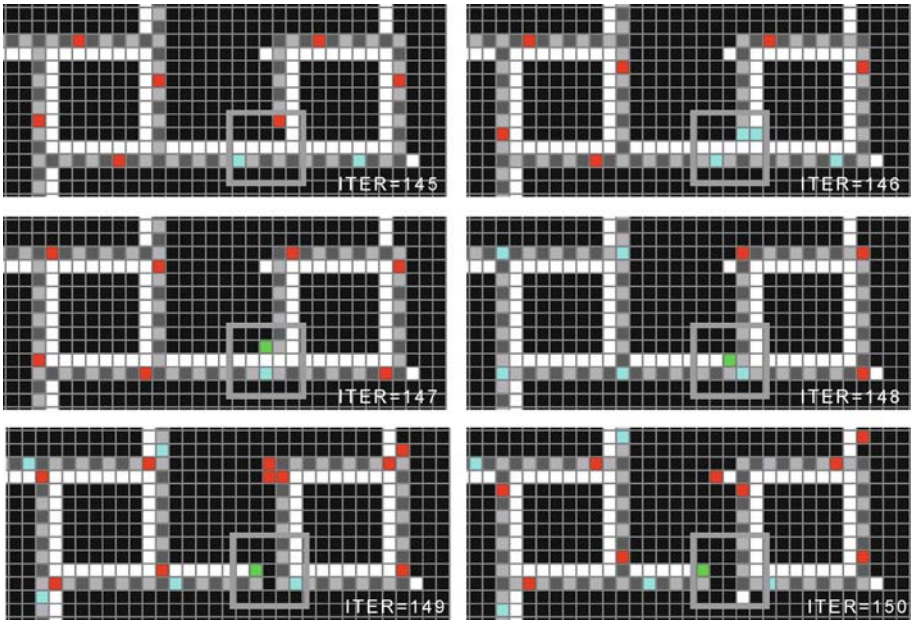
In the second part of the section, we will look at a set of approaches to self-replication that represent notable extensions to the approaches of von Neumann and Langton, because of different mechanisms (self-inspection), operat-



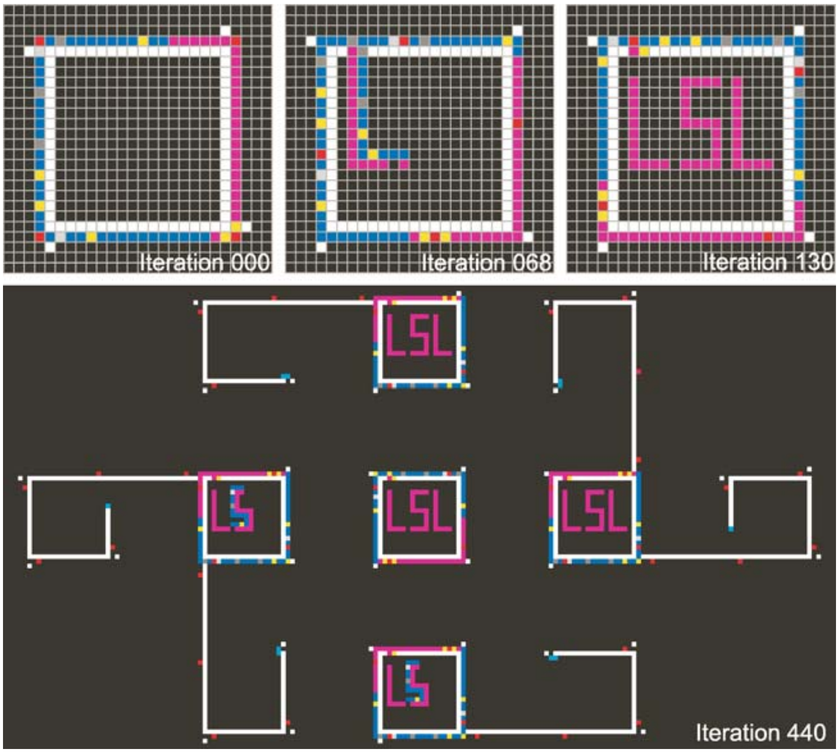
Self-Replication and Cellular Automata, Figure 12
The constructing arm begins to extend



Self-Replication and Cellular Automata, Figure 13
The new sheath has been fully constructed and a copy of the data is sent from the parent to the offspring



Self-Replication and Cellular Automata, Figure 14
The copy is complete and the constructing arm retracts



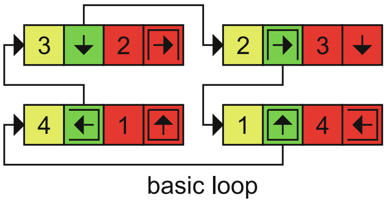
Self-Replication and Cellular Automata, Figure 15
The LSL automaton at different iterations

ing milieus (three-dimensional or self-timed CA), or design rules (evolutionary approaches).

Self-Replication in Hardware: The Tom Thumb Algorithm

In past years, we have devoted considerable effort to research on self-replication, studying this process from the point of view of the design of high-complexity multi-processor systems (with particular attention to next-generation technologies such as nanoelectronics). When considering self-replication in this context, Langton's loop and its successors share several weaknesses. Notably, besides the lack of functionality of Langton's loop (remedied only partially by its successors), which severely limits its usefulness for circuit design, each of these automata is characterized by a very loose utilization of the resources at its disposal: the majority of the elements in the cellular array remain in the quiescent state throughout the entire self-replication process.

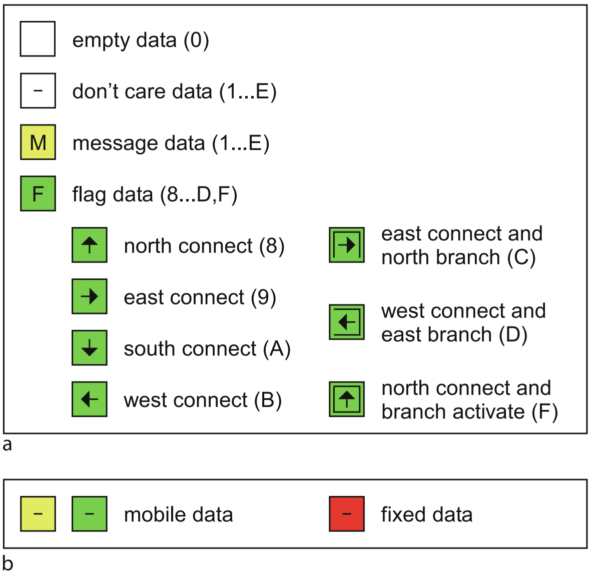
A new loop was then developed specifically to address these very practical issues. In fact, the system is targeted to the implementation of self-replication within the context of digital circuits realized with programmable logic devices (the states of the cellular automaton can then be seen as the configuration bits of the elements of the device). The new loop is based on an original algorithm, the so-called *Tom Thumb algorithm* [18,19].



Self-Replication and Cellular Automata, Figure 16
Tom Thumb algorithm: basic loop

The minimal loop compatible with this algorithm is made up of four cells, organized as a square of two rows by two columns (Fig. 16). Each cell is able to store in its four memory positions four hexadecimal characters of an artificial *genome* (defined as the information required for the construction of the loop). The whole loop thus embeds 16 such characters.

The original genome for the minimal loop is organized as another loop, the *basic loop*, of eight hexadecimal characters, i. e. half the number of characters in the minimal



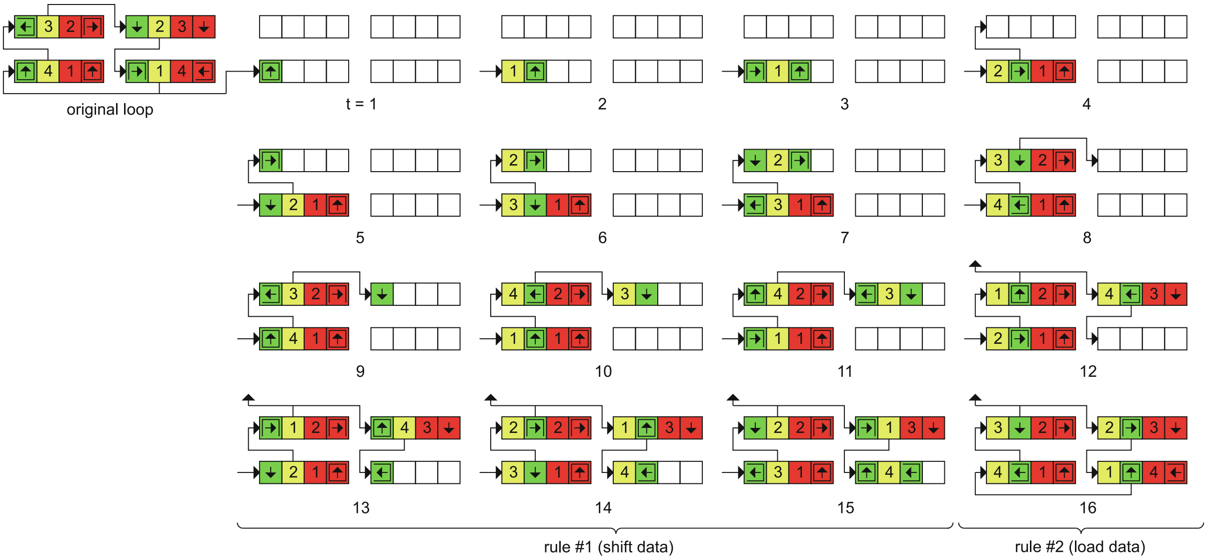
Self-Replication and Cellular Automata, Figure 17
a Graphical and hexadecimal representations of the 15 characters forming the alphabet of the artificial genome. b Graphical representation of the status of each character

loop, moving counterclockwise by one character at each time step.

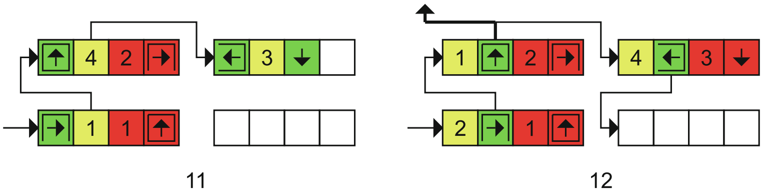
The 15 hexadecimal characters that compose the alphabet of the artificial genome are detailed in Fig. 17a. They are either *empty data* (0), *message data* ($M = 1 \dots E$), or *flag data* ($F = 8 \dots D, F$). Message data will be used to configure our final artificial organism, while flag data are indispensable for constructing the skeleton of the loop. Furthermore, each character is given a status and will eventually be *mobile data*, moving indefinitely around the loop, or *fixed data*, definitely trapped in a memory position of a cell (Fig. 17b). It is important to note that, while in this simple example the message data can take value from 1 to E, the Tom Thumb algorithm is perfectly scalable in this respect, that is, the size of the message data can be increased at will, while the flag data remain constant. This is a crucial observation in view of the exploitation of this algorithm in a programmable logic device, where the message data (the configuration data for the programmable elements of the circuit) are usually much more complex.

At each time step ($t = 1, 2, \dots$), a character of the *original loop* is sent to the lower leftmost cell (Fig. 18). The construction of the loop, i. e., storing the fixed data and defining the paths for mobile data, depends on two rules:

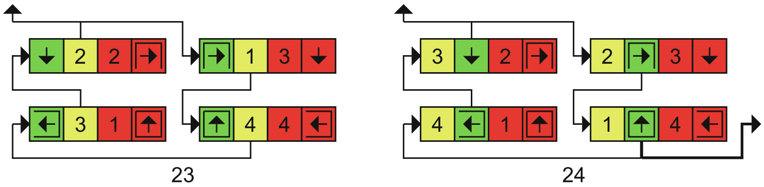
- If the four, three, or two rightmost memory positions of the cell are empty (blank squares), the characters are shifted by one position to the right (rule #1: shift data).
- If the rightmost memory position is empty, the characters are shifted by one position to the right (rule #2: load data). In this situation, the two rightmost characters are trapped in the cell (fixed data), and a new connection is established from the second leftmost position toward the northern, eastern, southern or western cell, depending on the fixed flag information (in Fig. 18, at



Self-Replication and Cellular Automata, Figure 18
Construction of a first specimen of the loop



Self-Replication and Cellular Automata, Figure 19
Creation of a new daughter loop to the north (rule #3)



Self-Replication and Cellular Automata, Figure 20
Creation of a new daughter loop to the east (rule #4)

time $t = 4$, the fixed flag F = F determines a northern connection).

At time $t = 16$, 16 characters, i. e., twice the contents of the basic loop, have been stored in the 16 memory positions of the loop (Fig. 18). Eight characters are fixed data, forming the *phenotype* of the final loop, and the eight remaining ones are mobile data, composing a copy of the original genome, i. e., the *genotype*. Both *interpretation* (the construction of the cell) and *copying* (the duplication of the genetic information) have been therefore achieved.

The fixed data trapped in the rightmost memory positions of each cell remind us of the pebbles left by Tom Thumb for memorizing his way in the famous children's story, an analogy that gives our algorithm its name.

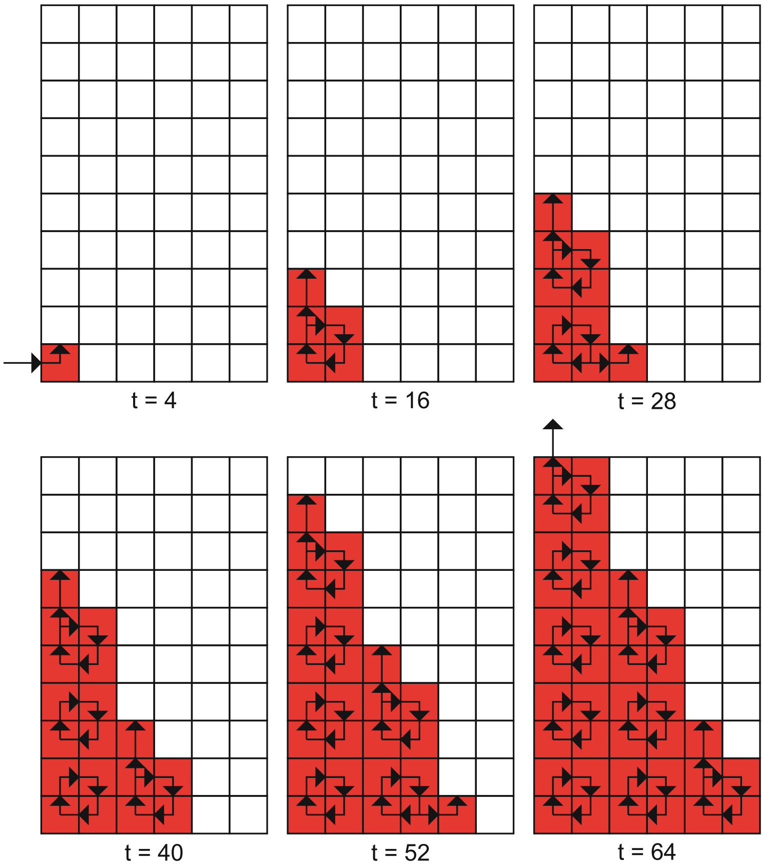
In order to grow loops in both horizontal and vertical directions, the mother loop should be able to trigger the construction of two daughter loops, northward and eastward. Two new rules are then necessary:

- At time $t = 11$ (Fig. 19), we observe a pattern of characters which is able to start the construction of the northward daughter loop; the upper leftmost cell is characterized by two specific flags, i. e., a fixed flag in the rightmost position, indicating a north branch (F = C) and the branch activation flag (F = F), in the leftmost position (rule #3: daughter loop to the north). The new path to the northward daughter loop will start from the second leftmost memory position ($t = 12$).
- At time $t = 23$ (Fig. 20), another particular pattern of characters starts the construction of the eastward daughter loop; the lower rightmost cell is characterized by two specific flags, i. e., a fixed flag indicating an east branch (F = D), in the rightmost position, and the branch activation flag (F = F), in the leftmost position (rule #4: daughter loop to the east). The new path to the eastward daughter loop starts from the second leftmost memory position ($t = 24$).

When two or more paths are activated simultaneously, a clear priority should be established between the different paths. Three growth patterns were chosen (Fig. 21), leading to four more rules:

- For loops in the lower row a collision occurs between the closing path, inside the loop, and the path entering the lower leftmost cell. The westward path has priority over the eastward path (rule #5).
- With the exception of the bottom loop, the inner path (i. e. the westward path) has priority over the northward path (rule #6) for the loops in the leftmost column.
- For all other loops, two types of collisions may occur: between the northward and eastward paths (2-signal collision) or between these two paths and a third one, the closing path (3-signal collision). In this case, the northward path will have priority over the eastward path (2-signal collision), and the westward path will have priority over the two other ones (3-signal collision)(rules #7 and #8).

We finally opted the following hierarchy: an east to west path has priority over a south to north path, which has priority over a west to east path.



Self-Replication and Cellular Automata, Figure 21
Growth of a colony of minimal loops represented at different time steps

The results of such a choice are as follows (Fig. 21): a closing loop has priority over all other outer paths, which makes the completed loop entirely independent of its neighbors, and the loops will grow bottom-up vertically. This choice is quite arbitrary and may be changed according to other specifications.

Unlike its predecessors, the Tom Thumb loop has been developed with a specific purpose beyond the theoretical study of self-replication. We believe that, in the not-so-distant future, circuits will reach densities such that conventional design techniques will become unwieldy. Should such an hypothesis be confirmed, self-replication could become an invaluable tool, allowing the engineer to design a single processing element, part of an extremely large array that would build itself through the duplication of the original element.

Current technology does not, of course, provide a level of complexity that would render this kind of process necessary. However, programmable logic devices (already among the densest circuits on the market) can be used as a first approximation of the kind of circuits that will become available in the future. Our loop is then targeted to the implementation of self-replication on this kind of device.

To this end, our loop introduces a number of features that are not present in any of the historical self-replicating loops we presented. Most notably, the structure of the loop (that is, the path used by the configuration data) is determined by the sequence of flags in the genome, implying that structures of almost any shape and size can be constructed and replicated using this algorithm, as long as the loop can be closed and that there is space for the daughter organisms. In practice, this implies that, if the Tom Thumb algorithm is used for the configuration logic of a programmable device, any of its configurations, and hence any digital circuit, can be made capable of self-replication.

In addition, particular care was given to develop a self-replication algorithm that is *efficient* (in the sense that it fully exploits the underlying medium, rather than leaving the vast majority of elements inert as past algorithms did), *scalable* (all the interactions between the elements are purely local, implying that organisms of any size can be implemented), and amenable to a *systematic design process*. These features are important requirements for the design of highly-complex systems based on either silicon or molecular-scale components.

Different Techniques and Environments

Von Neumann's and Langton's automata share a common basic technique to obtain self-replication: the construction of the new machine is directed through the *interpretation* of a description, coded as a sequence of states. In the case of von Neumann, this description (which, in biological terms, is usually identified as the genome of the artificial organism) is stored within the memory tape, which is read and interpreted by the universal constructor to build a copy of the machine. In Langton's case, the description is stored in the mobile data that runs within the sheath of the loop.

This mechanism of interpretation, while standard in many approaches, is not however unique: some examples of self-replicating CA exploit a different mechanism, that of *self-inspection*. In these approaches, instead of reading and interpreting a description, the self-replicating automaton inspects itself and produces a copy of what it finds. While less general than the universal constructor (obviously, the machine can only build an exact copy of itself), the functionality of this approach is similar to that of Langton's loop. Indeed, the most representative example of self-inspection is that of a self-replicating loop [12]. A more recent example is a variation of the Tom Thumb algorithm, where self-inspection was used to self-replicate a small processor within a field-programmable gate array [30].

And while the Tom Thumb algorithm targets in priority silicon-based circuits, other approaches have tried to explore alternative environments that, in some way, might more closely resemble the kind of technologies that will be available in the future. An example is Morita and Imai's study of self-replication in the context of reversible cellular automata [20] (in a reversible CA, every configuration has at most one predecessor), inspired by reversible logic in digital circuits.

Similarly, Peper et al. [24,38] have developed self-replicating structures in Self-Timed Cellular Automata (STCA). This kind of automata do not rely on a global synchronization mechanism to update the states of the cells, but rather the state transitions only occur when triggered by transitions in neighboring cells. The basic assumption in this work is that STCA is a model that might more closely resemble molecular-scale nanoelectronic devices.

A final example in this context is the three-dimensional extension of self-replication, usually based on the assumption that silicon, with its rigidly two-dimensional structure, will one day be superseded by a technology that can exploit all three dimensions. In this context, Imai et al. [13] have extended their reversible approach, with the assumption that reversible logic is more amenable to an extension to three dimensions than conventional logic because of the greatly reduced power dissipation. Stauffer at al. [36] have also shown that the Data and Signal Cellular Automaton (DSCA) approach, designed to simplify the implementation of CA in digital hardware, can be extended to realize self-replication in three dimensions.

The study of self-replicating CA in the context of new technologies holds the promise of one day bringing a major contribution to computation. To determine how self-replication might be useful in this context, some attempts have been made at using self-replicating structures for computation. An example of this approach is the work of Chou and Reggia [8], who use self-replication as a mechanism to obtain massively parallel machines which can potentially be used to solve hard problems (the example used in the paper is the NP-complete problem of satisfiability).

An attempt was also made to perform computation using Tempesti's loops. In alternative to embedding a complex program, this kind of loops are used to perform computation by inter-loop communication. Using the *collision-based computing* paradigm, Petraglio et al. [28] have shown that it is possible to implement arithmetic operations by passing messages from one loop to another after building a network structure through self-replication. This approach, while valuable from a theoretical standpoint, shares however the same weakness of other loop-based computing approaches in that the poor utilization of resources makes a physical realization of such a system highly impractical.

Another problem to be solved for a practical implementation of self-replicating structures is their design: few approaches have attempted to define a precise methodology to define and create self-replicating structures. In this context, several researchers have attempted to use evolutionary techniques to find automatically self-replicating machines. In this context, the work of Sayama et al. has gone through several iterations [31,32,33] in an attempt to define loops that evolve through the self-replication process towards "fitter" individuals. Chou and Reggia [7], on the other hand, use evolution to find novel self-replicating structures within a CA, whereas Pan and Reggia [23] studied the conditions in which self-replicating structures might spontaneously emerge in a cellular space.

Future Directions

Historically, self-replication in cellular automata began as a paradigm to achieve fault tolerance in computing devices. In the following decades, much of the emphasis shifted towards a more "theoretical" approach where self-replication was considered as part of a more general investigation into the application of biologically-inspired mechanisms to computing. And while the latter approach remains an active research topic, the original paradigm was somewhat set aside because technology would not allow a practical implementation of self-replication in digital hardware.

More recently, however, advances in electronic devices (notably with the introduction of programmable devices, or FPGAs), together with emerging technological issues have rekindled interest in self-replication in a context similar to von Neumann's original work. In particular, the drastic device shrinking, low power supply levels, and increasing operating speeds, which accompany the technological evolution of silicon to deeper submicron levels, significantly reduce the noise margins and increase the soft-error rates [42]. In addition, the nascent field of nanoelectronics holds great promise for the future of computing devices, but introduces extremely high fault rates (e. g., [25]) and complex layout issues.

Thus, self-replication is currently attracting a considerable amount of attention for the same reasons that initially pushed von Neumann to investigate it as a possible solution to reliability and layout issues. Fault tolerance and self-organization are thus becoming the focal point of research in the field and the features of molecular-scale electronics seem to imply that self-replication at the device level will be an extremely useful paradigm in next-generation devices.

Bibliography

1. Asprey W (1992) John von Neumann and the Origins of Modern Computing. The MIT Press, Cambridge
2. Banks ER (1970) Universality in Cellular Automata. In: Proc. IEEE 11th Annual Symposium on Switching and Automata Theory, Santa Monica, CA, pp 194-215
3. Beuchat J-L, Haenni J-O (2000) Von Neumann's 29-State Cellular Automaton: A Hardware Implementation. IEEE Trans Education 43(3):300–308
4. Buckley WR, Mukherjee A (2005) Constructibility of Signal-Crossing Solutions in von Neumann 29-State Cellular Automata. Proc. 2005 Int. Conf. on Computational Science. LNCS, vol 3515. Springer, Berlin, pp 395–403
5. Burks A (ed) (1970) Essays on Cellular Automata. University of Illinois Press, Urbana
6. Byl J (1989) Self-Reproduction in Small Cellular Automata. Physica D 34:295–299
7. Chou H-H, Reggia JA (1997) Emergence of self-replicating structures in a cellular automata space. Physica D 110(3-4):252–276
8. Chou H-H, Reggia JA (1998) Problem solving during artificial selection of self-replicating loops. Physica D 115(3-4):293–312
9. Codd EF (1968) Cellular Automata. Academic Press, New York
10. Freitas RA Jr, Gilbreath WP (eds) (1980) Advanced Automation for Space Missions. In: Proc. 1980 NASA/ASEE Summer Study, Scientific and Technical Information Branch (available from U.S. G.P.O.), Washington, DC
11. Freitas RA Jr, Merkle RC (2004) Kinematic Self-Replicating Machines. Landes Bioscience, Georgetown
12. Ibanez J, Anabitarte D, Azpeitia I, Barrera O, Barrutieta A, Blanco H, Echarte F (1995) Self-inspection based reproduction in cellular automata. In: Proc. 3rd European Conf. on Artificial Life (ECAL95). LNCS, vol 929. Springer, Berlin, pp 564–576
13. Imai K, Hori T, Morita K (2002) Self-reproduction in three-dimensional reversible cellular space. Artif Life 8(2):155–174
14. Langton CG (1984) Self-Reproduction in Cellular Automata. Physica D 10:135–144
15. Lee C (1968) Synthesis of a Cellular Computer. In: Applied Automata Theory. Academic Press, London, pp 217–234
16. Lohn JD, Reggia JA (1997) Automatic Discovery of Self-Replicating Structures in Cellular Automata. IEEE Trans Evol Comput 1(3):165–178
17. Mange D, Sipper M, Stauffer A, Tempesti G (2000) Towards Robust Integrated Circuits: The Embryonics Approach. Proc IEEE 88(4):516–541
18. Mange D, Stauffer A, Petraglio E, Tempesti G (2004) Self-replicating loop with universal construction. Physica D 191:178–192
19. Mange D, Stauffer A, Peparolo L, Tempesti G (2004) A Macroscopic View of Self-Replication. Proc IEEE 92(12):1929–1945
20. Morita K, Imai K (1996) Self-reproduction in a reversible cellular space. Theor Comput Sci 168:337–366
21. Nehaniv CL (2002) Self-Reproduction in Asynchronous Cellular Automata. In: Proc. 2002 NASA/DoD Conf. on Evolvable Hardware (EH02). IEEE Computer Society, Washington, DC, pp 201–209
22. Nourai F, Kashef RS (1975) A Universal Four-State Cellular Computer. IEEE Trans Comput 24(8):766–776
23. Pan Z, Reggia J (2005) Evolutionary Discovery of Arbitrary Self-replicating Structures. In: Proc. 5th Int. Conf. on Computational Science (ICCS 2005) - Part II. LNCS, vol 3515. Springer, Berlin, pp 404–411
24. Peper F, Isokawa T, Kouda N, Matsui N (2002) Self-Timed Cellular Automata and their computational ability. Future Gener Comput Syst 18(7):893–904
25. Peper F, Lee J, Abo F, Isokawa T, Adaki S, Matsui N, Mashiko S (2004) Fault-Tolerance in Nanocomputers: a Cellular Array Approach. IEEE Trans Nanotechnol 3(1):187–201
26. Perrier J-Y, Sipper M, Zahnd J (1996) Toward a Viable, Self-Reproducing Universal Computer. Physica D 97:335–352
27. Pesavento U (1995) An Implementation of von Neumann's Self-Reproducing Machine. Artif Life 2(4):337–354
28. Petraglio E, Tempesti G, Henry J-M (2002) Arithmetic Operations with Self-Replicating Loops. In: Adamatsky A (ed) Collision-Based Computing. Springer-Verlag, London, pp 469-490
29. Reggia JA, Armentrout SA, Chou H-H, Peng Y (1993) Simple Systems That Exhibit Self-Directed Replication. Science 259:1282–1287

30. Rossier J, Thoma Y, Mudry PA, Tempesti G (2006) MOVE Processors that Self-Replicate and Differentiate. In: Proc. 2nd Int. Workshop on Biologically-Inspired Approaches to Advanced Information Technology (Bio-ADIT06). LNCS, vol 3853. Springer, Berlin, pp 328–343
31. Salzberg C, Antony A, Sayama H (2004) Evolutionary dynamics of cellular automata-based self-replicators in hostile environments. BioSystems 78:119–134
32. Sayama H (1998) Introduction of Structural Dissolution into Langton's Self-Reproducing Loop. In: Artificial Life VI, Proc. 6th Int. Conf. on Artificial Life, MIT Press, Cambridge, pp 114–122
33. Sayama H (2000) Self-replicating worms that increase structural complexity through gene transmission. In: Artificial Life VII: Proc. 7th Int. Conf. on Artificial Life, MIT Press, Cambridge, pp 21–30
34. Sipper M (1995) Studying artificial life using a simple, general cellular model. Artif Life 2(1):1–35
35. Sipper M, Mange D, Stauffer A (1997) Ontogenetic Hardware. BioSystems 44:193–207
36. Stauffer A, Mange D, Petraglio E, Vannel F (2004) DSCA Implementation of 3D Self-Replicating Structures. In: Proc. 6th Int. Conf. on Cellular Automata for Research and Industry (ACRI2004). LNCS, vol 3305. Springer, Berlin, pp 698–708
37. Stauffer A, Sipper M (2004) The Data-and-Signals Cellular Automaton and Its Application to Growing Structures. Artif Life 10(4):463–477
38. Takada Y, Isokawa T, Peper F, Matsui N (2006) Universal Construction and Self-Reproduction on Self-Timed Cellular Automata. Int J Mod Phys C 17(7):985–1007
39. Tempesti G (1995) A New Self-Reproducing Cellular Automaton Capable of Construction and Computation. In: Proc. 3rd European Conf. on Artificial Life. LNAI, vol 929. Springer, Berlin, pp 555–563
40. Tempesti G (1998) A Self-Repairing Multiplexer-Based FPGA Inspired by Biological Processes. Ph.D. Thesis, Ecole Polytechnique Fédérale de Lausanne (EPFL)
41. Trimberger S (ed) (1994) Field-Programmable Gate Array Technology. Kluwer, Boston
42. Various (1999) A D&T Roundtable: Online Test. IEEE Design Test Comput 16(1):80–86
43. Vitanyi PMB (1973) Sexually reproducing cellular automata. Math Biosci 18:23–54
44. Burks AW (ed) von Neumann J (1966) The Theory of Self-Reproducing Automata. University of Illinois Press, Urbana

Semantic Web

WENDY HALL, KIERON O'HARA
Intelligence, Agents, Multimedia Group, School
of Electronics and Computer Science, University
of Southampton, Southampton, United Kingdom

Article Outline

Glossary

Dereferencing Dereferencing a URI is using the URI to identify the object or resource it refers to.

Folksonomy 'Folksonomy' is a neologism applied to structures that emerge from the practice of 'tagging' Web content. In some Web 2.0 applications, users can apply a tag (a descriptive term) to content such as a photograph or video clip. The tags need only be meaningful to the individual tagger, but if a large enough number of users tag content, descriptive structures analogous to more formal ontologies can emerge that are meaningful to wide communities.

GRDDL Gleaning Resource Descriptions from Dialects of Languages (GRDDL, pronounced 'griddle') is a mechanism for helping bootstrap the Semantic Web, by extracting RDF from XML documents, using transformations expressed in XSLT. GRDDL became a W3C recommendation in 2007.

Metadata Metadata is data about data. In the context of the Semantic Web, metadata are also called 'markup' or 'annotations'. Because one aim of the Semantic Web is to support machine processing of information, metadata are helpful in describing the content of data. For example, metadata attached to a series of numerals could explain that it represents a zip code or a height or a population figure.

Ontology An ontology defines, describes and constrains the concepts and relationships that are used in some particular domain of knowledge. Ontologies may therefore have an important role in data sharing, for example by providing a means of expressing which concepts (in the ontology) are referred to by particular terms (in a set of databases, which may use widely differing vocabularies).

OWL The Web Ontology Language (OWL) is a language for describing and sharing ontologies on the Web. It is an extension of RDF, and has three variants. OWL Full is maximally expressive and compatible with RDF (any legal RDF document is a legal OWL Full document), but is undecidable. OWL DL is based on a series of restrictions of OWL Full which support efficient reasoning, at the cost of losing the strong connection with RDF (not all RDF documents are legal OWL DL documents). OWL Lite is even more restricted in expressivity, but is easier to grasp. OWL became a W3C recommendation in 2004.

RDF The Resource Description Framework (RDF) is a standard framework for representing data on the Web, representing it in a three-place relation, of subject-relation-object form, called a *triple*. It uses URIs to refer not only to the two items related, but also to the relationship asserted between them. In this way it extends the linking structure of the Web by allowing the nature of the link asserted to be described. Its syntax is XML-based, to support syntactic interoperability between the two. RDF became a W3C recommendation in 1999.

RDF(S) RDF Schema (RDF(S) or RDFS) is a language for representing information on the Web. It is an extension of RDF to allow the description of the relationships which can be asserted between resources using RDF. So, for instance, RDF allows the assertion of the relationship 'author' between, say, 'Herman_Melville' and 'Moby_Dick', but RDF(S) is required to assert the properties of the 'author' relationship (e. g. that every document has at least one author, or that the 'author' relationship is the inverse of the 'written_by' relationship). RDF(S) became a W3C recommendation in 2004.

Rules and RIF Rules govern the transformation of, and inference from, data. In particular, when data are being shared, it may be useful to make basic inferences over the data (for example, to determine whether two names refer to the same object or different objects). Rule-based knowledge such as this cannot be expressed in individual data stores, and may be hard to express in an ontology. The Rule Interchange Format (RIF) is a language, not complete at the time of writing, to express the most common or basic types of rule.

SPARQL SPARQL is a special query language designed specifically for querying data stored in RDF ('SPARQL' is a recursive acronym standing for 'SPARQL Protocol And RDF Query Language', and is pronounced 'sparkle'). SPARQL became a W3C recommendation in January 2008.

Triples and triplestores A triple is a statement in RDF, consisting of a subject, an object, and a binary predicate that relates them. Each item of the triple is identified by a URI reference (or can be a string literal, such as a date or a number or name). A knowledge repository which contains RDF triples is called a triplestore. Triplestores may need to contain several millions of triples, and so need to be able to support fast querying at these potentially very large scales.

URI A Uniform Resource Identifier (URI) is a string of characters for identifying an abstract or physical object or resource. There are different URI schemes (i. e. different ways to restrict the syntax of a URI to make it meaningful). A Uniform Resource Locator (URL) is a particular type of URI that identifies its object by means of its access mechanism or 'location' in the network. URIs are important in that they act as a standard way of referring to objects on the Web.

W3C The World Wide Web Consortium (W3C) is an international non-profit consortium which coordinates the development of Web standards, founded in 1994 under the directorship of Sir Tim Berners-Lee. The W3C groups together all the bodies involved in specifying Semantic Web standards into the Semantic Web Activity. A W3C working group works on a standard for a particular formalism, and when the standard is judged by the working group to be in a final form, fit for purpose and properly interoperable with other W3C standards, it ratifies the formalism by recommending it. Hence a W3C recommendation is an important standard.

Web of data The Web of Data is another way of referring to or explaining the vision of the Semantic Web, which emphasises the idea of creating links between data rather than documents (as on the current World Wide Web). Linking data, as with linking documents, enables them to be reused in interesting or unexpected contexts (when the data are interpreted using explicit semantic theories of the sort provided by ontologies). RDF is the anticipated mechanism for creating the links between data; dereferencing one or more of the URIs in an RDF triple will lead to descriptions of the resources referred to, which in turn are likely to contain further triples, which can again lead to dereferencing and so on.

Web science Web Science is the multidisciplinary activity of trying to understand the two-way dynamic relationship between Web technology and wider society, in order to ensure that the technological changes made to the Web (including the Semantic Web) are generally beneficial rather than otherwise.

XML The eXtensible Markup Language (XML) is a language for allowing users to tag or mark up content using tags of their own devising (for instance based on a particular vocabulary used in a small community of practice). As such, XML can serve as a basic data exchange format between applications. It is an advance on other well-known markup languages (such as HTML, a standard language for marking up content for display on the Web) in that it separates instructions to do with content and document structure from those to do with formatting. It is a basic language for representing and exchanging structured information.

Definition of the Subject

The Semantic Web is a proposed extension to the World Wide Web (WWW) that aims to provide a common framework for sharing and reusing data across applications. The most common interfaces to the World Wide Web present it as a Web of Documents, linked in various ways including hyperlinks. But from the data point of view, each document is a black box – the data are not given independently of their representation in the document. This reduces its power, and also (as most information needs to be extracted from documents by a human agent) inhibits the use of automatic information processing methods on the Web. The Semantic Web is an effort, steered by the World Wide Web Consortium, to develop a set of protocols, formalisms and standards to transform the Web into a *Web of Data*. Links would be between data, and data could be accessed independently of the applications that created them. This would allow both the sharing of data, and the amalgamation of data from different sources, using heterogeneous formats, in new contexts.

Introduction

The idea of the Semantic Web (SW), of exploiting the possibilities for serendipitous reuse of linked data, dates back at least to Sir Tim Berners-Lee's plenary talk at the first International World Wide Web Conference at CERN in Geneva in 1994 [33]. In that talk, Berners-Lee argued that there is too little machine-readable information on the WWW as was currently constituted. "The meaning of the documents is clear to those with a grasp of (normally) English, and the significance of the links is only evident from the context around the anchor. To a computer, then, the Web is a flat, boring world devoid of meaning. This is a pity, as in fact documents on the Web describe real objects and imaginary concepts, and give particular relationships between them. … Adding semantics to the Web involves two things: allowing documents which have information in machine-readable forms, and allowing links to be created with relationship values. Only when we have this extra level of semantics will we be able to use computer power to help us exploit the information to a greater extent than our own reading." Indeed, the original vision of the WWW was intended to support greater machine understanding of people's work and interactions; the 'flat' understanding of the world that the WWW produces in machines is a first step towards the richer vision.

The SW was from the beginning conceived as a set of layered standards and formalisms (see Sect. "The Layered Model of the Semantic Web" for details). The development of the Resource Description Framework (RDF) in the 1990s was a key technology [80]. RDF is an important formalism, as it allows expression not only *of* the link between two objects, but also about the *nature* of the link itself. Hence, one can follow a chain of links not only via the objects linked, but also the types of links involved. In the WWW of documents, links connect documents written in the Hypertext Markup Language HTML; in the SW, links in RDF connect not only documents but arbitrary things (objects and relationships) identified by the Uniform Resource Identifiers (URIs [38]) in the RDF triples representing the data.

The ability to move between data linked in such a way opens up the possibility of exposing data to the Web, and then being able to access such data from any application. As a simple example, consider personal information such as one's bank statements, information-based resources such as digital photographs, and an application such as a calendar or diary. Each of these depends on data which is controlled by the applications that use them. But in a genuine Web of Data, we could link these data in a productive way – something as simple as being able to present one's financial information in one's calendar. The metadata in one's photographs often includes information about the time of their creation; an application able to get at the photograph metadata and the data on one's calendar might be able to suggest where the photo was taken, and its possible location. The ability to use all these data in a constructive way is impossible without a Web of linked data to enable applications to move between the data sources.

In this way, the SW changes our model of the value of information. Currently, it is generally presumed that the value of information stems from its *scarcity* – people and organizations gain value from information they have gathered, and are given monopoly rights to exploit that information via such legal contrivances as copyright, intellectual property rights, licensing, and so on. Even when organizations do not resort to the law, they will make great investments in protecting trade secrets. However, this scarcity-based model seems inadequate for the digital age.

In the first place, as economist William Baumol has argued, the social benefits from unlicensed use of 'protected' knowledge and innovation, were already large in the predigital economy, and indeed account for much of our wealth today: "some 80 percent of the benefits [innovation] may – i. e. change round brackets to square, as this is an editorial interpolation by us plausibly have gone to persons who made no direct contribution to innovation. The rather startling implication of all this is that the spillovers of innovation, both direct and indirect, can be estimated to constitute well over half of current [US] GDP – and it can even be argued that this is a very conservative figure" (see p. 135 in [32]). And secondly, the Internet and the Web have made it harder to preserve monopoly rights to information, as copying and distribution reduces the marginal cost to producers to close to zero. Although many media companies have taken rearguard action to protect their intellectual property, so simple is the distribution model on the Web that the basis of the value of information is rapidly switching from scarcity to *abundance*. It is the large quantity of data, that can be placed in novel and unintended contexts with little cost, that makes it increasingly valuable in the age of digital technologies – and it is this abundance that the SW is designed to foster.

One of the major drivers of the SW has been the transformation of science into *e-science*, a computer-enabled, data-heavy view of science as the analysis of the very large quantities of information that improved instrumentation, larger computer power, more prevalent sensor networks and greater memory storage have released. Several disciplines have seized on the opportunity to exploit such data, which are often available only in diverse and heterogeneous datasets. In particular, interdisciplinary research is growing in importance, requiring data developed in different disciplines, using a confusion of vocabularies and methods of collection. Methods for dealing with such large and heterogeneous datasets are required in many areas, including the life sciences, climate research, medicine and epidemiology, and genomics, to name but four, which explains the interest in many of these fields in the SW.

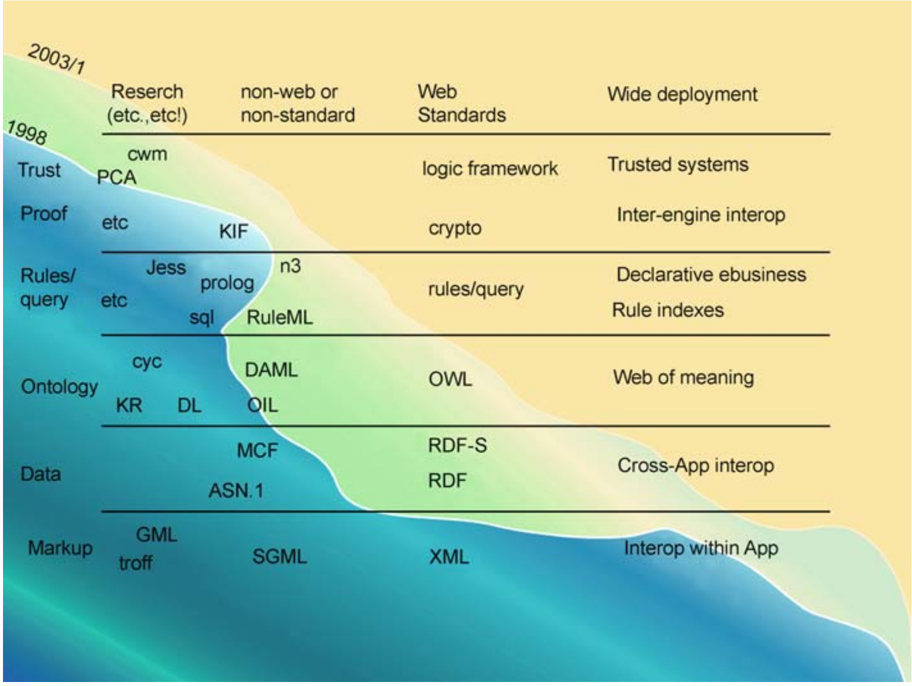
The use of the SW in such large, public projects was perhaps predictable, but much debate and discussion has focused more on the individual's interface to it. We discuss this in more detail below, but one reason for this was that the landmark publication for the public view of the SW, an article in *Scientific American* for 2001, written by Berners-Lee, James Hendler and Ora Lassila [37], developed the idea of a Web of Data with a number of household gadgets interfacing with it. The possibilities envisaged in their scenario included: a telephone that turned down the volume of all local devices with volume controls when it rang; an agent that could plan a program of medical care; and a calendar that could integrate this information to adjust a set of appointments. The point of the article was not the impressive set of agents, but rather the Web of Data that sat underneath them. However, many readers focused on the gadgets, and – given that such gadgets are not at the time of writing very common or effective – have concluded either that the SW has been a failure, or that it was an unrealistic vision from the beginning (see Sect. "Controversies").

In 2006, a further publication by Berners-Lee, together with Nigel Shadbolt and Wendy Hall, appeared in the publication *IEEE Intelligent Systems* de-emphasising the agents and focusing on the idea of the SW as a Web of Data or actionable information [101]. This paper argued that the agents described in 2001 could only flourish when standards for data sharing are well-established. The need for such standards, and for SW technologies in general, was growing, thanks to developments such as e-science, information-based medicine, and e-government.

Since the late 1990s, the World Wide Web Consortium (W3C), under the direction of Berners-Lee, has led the drive to create the standards. Figure 1 shows a diagrammatic representation of the progress of the SW in the development of its layered standards (the hierarchically-arranged layers are marked in Fig. 1 down the left hand column as markup/data/ontology etc.). The development of each layer goes through a long, sometimes tortuous, process of research and discussion. Initial stages of research, sometimes competitive, after some time produce a rough consensus about the general properties of a formalism to implement a layer. At that point, consideration is undertaken about creating a Web standard by the W3C.

To create a standard, a group, representative of stakeholders from academe and private enterprise, is assembled by the W3C, which first creates a *working draft*, which is released under that status for review by the community. Commentary is welcomed, and the draft may be changed dramatically. Once the group responsible for the standard is satisfied that it is fully capable of doing what it required, it is released as a *candidate recommendation*, when it is critiqued in terms of the practicability of its implementation. The next stage is to become a *proposed recommendation*, when it is submitted to the W3C advisory council. Finally, it is released as a W3C recommendation. It generally takes years to negotiate these various stages. Currently, the markup language XML, the data representation and interchange languages RDF and RDF(S), and the ontology language OWL are full recommendations. The next layer up from the ontology layer is that of rules and querying: the query language SPARQL became a W3C recommendation in January 2008 [96], while the rule expression language RIF is at the time of writing at the working draft stage [41].

Figure 1 represents the historical progress of the SW in terms of a 'wave' rolling over 'dry land'. The depth of the 'sea' indicates the extent to which SW standards have been accepted and been deployed widely. The SW wave is traveling from the bottom left of the diagram to the top



Semantic Web, Figure 1
The wave of development of the Semantic Web [51]

right; the net result of this is that the lower levels of the SW are coming into being and wide acceptance ahead of the higher levels. RDF is, at the time of writing, reasonably widely deployed; SPARQL has recently become a recommendation; trust (the highest level represented here) is still a research issue, and the consensus about what form a standard to represent and promote trust in data is still in the process of formation.

Linking Data

The underlying aim of the SW is to allow data to be explored and queried on the Web, analogously to the way that documents are currently investigated online. One precondition for this is obviously the publishing of data on the Web, but another is to create the links that allow data to be explored. RDF allows representation of data in such a way that anything referred to in the data can be linked to, and from. If URIs are used to name things, common naming schemes are allowed to emerge; one of the most important is the Hypertext Transfer Protocol (HTTP [59]), which affords a straightforward mechanism for people to look up the names. In a properly linked Web of Data, the URI, once looked up, should provide access to useful information about the resource named, as well as useful links out to other data.

Links can be made using various mechanisms, the simplest of which is to use a URI that points to another. For example (taken from [35]), someone might describe some relationships in RDF as follows:

```
<rdf:Description about="#albert"
   <fam:child rdf:Resource="#brian">
   <fam:child rdf:Resource="#carol">
</rdf:Description>
```

This RDF is about three resources given the local identifiers '#albert', '#brian' and '#carol', and might be placed in a file called '<http://example.org/smith>'. The architecture of the Web can use these names to provide a global identifier for the three resources; for instance "http://example.org/smith#albert" refers to #albert, and so on. And now there is a global identifier, links can be made. For instance, a document '<http://example.org/jones>' might contain the following RDF:

```
<rdf:Description about="#denise"
   <fam:child rdf:Resource="#edwin">
   <fam:child rdf:Resource=
                        "http://example.org/smith#carol">
</rdf:Description>
```

Here a series of relationships between resources #denise, #edwin and #carol have been asserted, but the datum about #carol makes it possible to link to the data in the other file. Someone following the link dereferences the URI, i. e. decomposes 'http://example.org/smith#carol' into two parts: the part before the '#' which gives the name and location of the file; and '#carol' which is the local identifier in that file. Hence the information about #carol in the first file can be accessed thanks to the link included in the second file. The series of links between different resources can be represented, at least on a small scale, graphically, as in Fig. 3. This is the simplest way of linking data, though there are others [35]. And if the URI used for reference is created under a widely-supported system in a community, then the prospects for linking data are that much larger.

One of the main drivers of the SW is the vast quantity of data available in the form of relational databases (RDBs) (RDBs), which often exist in isolation from each other. Each database has its own value, but as argued above, the major source of informational value in the digital age is abundance, the possibilities for serendipitous reuse of data by placing it in fruitful contexts. To that end, a key aim of the SW is to harvest the large amount of data held in RDBs – a much larger quantity than is currently available in the document Web – and to support its amalgamation. The net result will be to facilitate the treatment of all the data as, in effect, sitting in a single queryable database.

Much of the current Web is supported by larger databases that sit below the level of what can be seen on a webpage; these databases are known as the *deep Web*, as opposed to the *shallow Web* on webpages. So, for instance, when one looks at one's bank statement on the shallow Web, the webpage that the bank's site creates uses a much larger quantity of data about one's bank account than is visible at any one time. Hence an alternative way of looking at this role for the SW is as a method for allowing users to query the whole of the deep Web, rather than simply the information released onto the shallow Web by the current set of implemented applications.

An RDB consists of a series of *tables*, which consist of *rows* or *records* of individual data items. Each record consists of a set of values of *fields* or *attributes*. The records (rows) and fields (columns) together can be conceived as a table or matrix (m records and n fields give us an $m \times n$ matrix) [50]. So, for instance, each record may represent an individual person, while a particular field may represent zip codes. The value placed in the zip code field of a person's record is therefore the value of that person's zip code. This tabular structure for RDBs can be mapped straightforwardly into an RDF representation. The record can be seen as the subject of an RDF triple, the field name is a link or property type, while the value of that field is the object of the triple. Hence the person who is represented by the record in the example RDB above would be represented by the first item of the triple; the zip code field would be represented by a zip_code property the second item of the triple; and the value of the zip code would be the third item. In that way, each cell of the RDB matrix is represented by an individual RDF triple, and the total set of triples would represent the entire database [34].

Having said that, there are additional factors about RDBs that are harder to capture in the RDF, and there are many open questions about the export of RDB. For instance, it may be that the database is definitive – that is, the institution holding the data has some responsibility for the data. It may be that the State of Texas holds a database of all Texas vehicle registration numbers; any car not on the database is not, as a matter of definition, registered in Texas. On a smaller scale, some RDBs allow a primary field for a unique identifier for the record, which also holds a significance beyond the particular piece of data. There are various ways of modeling these context-based properties of RDBs, perhaps most likely devolving the representation of such matters to the applications that use the data, via the ontologies, rules or query types that they use. See [34] for a worked example of methods to expose an RDB to the Web.

The process of exposing databases to the Web should not be too prescriptive – the whole point of the Web is as a decentralized collection of linked information, whether in the form of data or documents. The links are entirely democratic, and can be made between any pair of data items, or any pair of documents. This is where the power of the Web's ability to promote serendipitous reuse comes in. Attempting to fix the methods or languages used, or to 'police' the links made, will blur this vision, and create bottlenecks impeding information flow. Indeed, the widespread use of the Web is largely down to its non-prescriptive nature – prescription will simply drive users, who will not want particular information management strategies forced upon them, from the Web.

The result is a particularly untidy situation, unusual in the history of information management. If data, information or knowledge is generated within a single organization or affiliation (the usual situation for information managers before the growth of the Web), then information systems can trade on a number of simplifying assumptions. The size of such repositories can be assumed to be small or medium, and representation schemes would be planned and homogeneous. The quality of information would be likely to be high, and managers' trust in it correspondingly high. But on the Web scale, these assumptions fail. The amount of data available for query may be extremely high, and represented in highly heterogeneous ways – rarely in the optimal way for the manager's task in hand. Information quality, and trust in that quality, would be very variable. Linking data using common URI schemes will never work perfectly, precisely because there are deliberately no enforcement mechanisms on the Web, and people cannot be forced to use any particular naming convention.

It is primarily for this reason that ontologies have always played a central role in the vision of the SW [57]. It is the ontology that puts the 'semantic' into 'Semantic Web'. Ontologies specify the vocabulary, concepts and relationships of a domain. They must be a rationalization of current practice, and managed and endorsed by a community. They act as a specification of the terms used in discussion. But they should not be too prescriptive – terms change over time, others are in constant dispute. Ontologies need to develop as a discipline or domain develops. Different areas will have different requirements of ontologies; some sciences will have large, publicly-managed ontologies which act as a public vocabulary standard, while others will make do with small, lightweight ontologies that only define the relationships between a few terms. The ontological requirements of any individual application may be quite small; the question for the application developer is whether to develop a small special-purpose ontology for her own individual purposes, or alternatively whether to reuse a larger, better-known ontology that may overspecify vocabulary for her purposes. Much will depend on the usual practices of her wider community. It should be noted that the SW project does *not* require a single overarching ontology, referring to and prescribing everything.

The ontology is key to being able to deal with heterogeneous datasets as described above; the data, and the terms used in it, must be mapped onto other terms held in common. Once this has been done, then databases can be understood in common terms – and, most crucially, the information they hold amalgamated and processed by machines. It is of course obvious to a human user that if one database of people has a field *ZC*, while another has a field *zip_code*, that there is at least a good chance that the two fields refer to the same attribute of people, viz., the zip codes of their addresses; a computer will merely try, and fail, to match the strings identifying each field. But if the computer is referred to an ontology and given mappings from the terms used in the two databases to the ontology's terms, then it can be told about the equivalence, and accordingly its inference space is opened up (for instance, it could make some inferences from the fact that $ZC(X) = zip\_code(y)$). Machine processing of this heterogeneous, distributed data is made possible by ontologies.

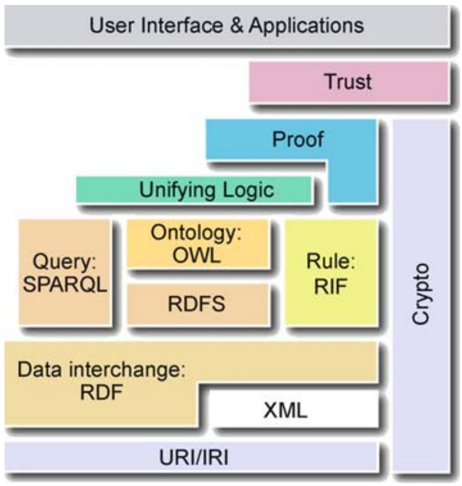
There are, of course, several issues pertaining to the use of ontologies in the SW, some of which will be discussed in greater detail in the Sect. "Controversies". The development of an ontology for an application is a (potentially large) initial cost to an application developer, and it is likely that ontologies, especially well-known ones, will be reused. To that end, searching for ontologies is likely to be a growth area in the future; there are already specialized search engines, such as Swoogle, dedicated to this task [8,54]. It may also be that one application might use several ontologies (for instance, an interdisciplinary scientific application may well reuse well-known ontologies from each discipline that it crosses). In that case, mappings *between* the ontologies will be important, and such mappings, as opposed to the ontologies themselves, could become the semantic basis for the application [78].

Building an ontology from scratch is always an option, especially if the application requires only relatively lightweight ontological apparatus [89]; again, special-purpose tools, such as Protégé are already available and well-used in the SW community [6,90]. Generating an ontology from an RDB can be done semi-automatically, and then mappings (which will be fairly straightforward, given the method of ontology generation) defined between the database and the ontology [105]. The problem is more complex if the aim is to map a legacy database onto an existing ontology. In particular, the mappings between the database and the ontology can be expected to be quite complex, and therefore very expressive languages will be required to describe them, such as the language R_2O [31].

The Layered Model of the Semantic Web

The Web, as a decentralized construction, cannot be created by fiat or prescription, which would limit its growth and create bottlenecks for information mobility. But to allow the Web of Data to reach fruition at the scale envisaged, several related tasks are required to be performed [101]. As discussed above, the W3C has devoted resources over the last few years to developing formalisms and standards to allow these tasks to be addressed, and the tasks themselves have also been arranged in a series of layers, depicted in a hierarchical diagram. The diagram has evolved with the vision of the SW, but is not dissimilar to its first incarnation, and at the time of writing is seen as in Fig. 2.

As noted above, and in Fig. 1, the development of these layers has been bottom-up, concentrating on the lower levels before work begins on those further up. The lower layers are now often the subject of W3C recommendations, while work on the upper layers remains generally more theoretical, and contains more open research questions.



Semantic Web, Figure 2
The layered view of the Semantic Web [39]

At the lowest level we have URIs (and IRIs, Internationalized Resource Identifiers, which are generalizations of URIs allowing non-ASCII characters to be used [56]). URIs identify resources in a global way – in other words, they are interpreted consistently across applications, unlike individual naming conventions, and therefore are central to the vision of a Web of Data [38]. Using a URI to identify a resource (whether that resource be a piece of information, a real-world object, an abstract concept, etc.) allows others to use the same identifier to link to the resource, refer to it, or retrieve a representation of it; this shifts the emphasis online from documents to data, and allows direct machine processing of data. If the URI scheme used is the Hypertext Transfer Protocol (http), then that is particularly helpful as http guides the user as to the location of the resource (although there are several other URI schemes, and indeed users can invent their own).

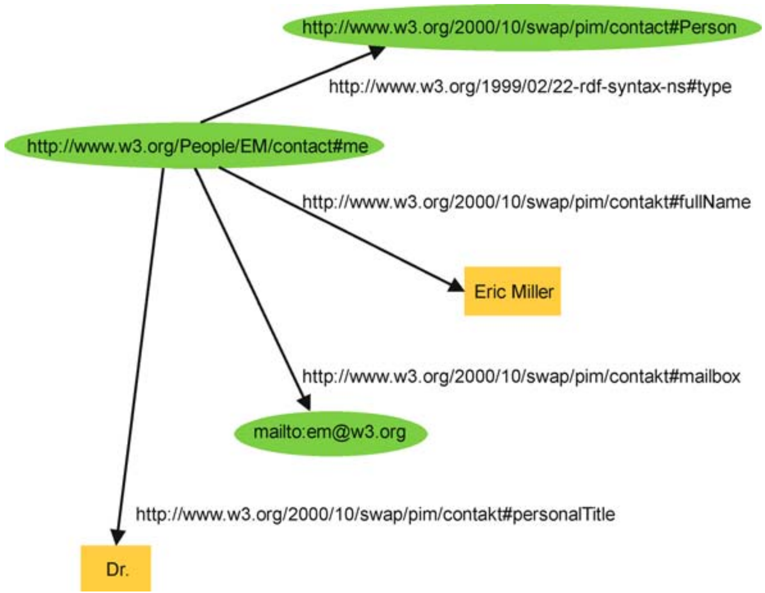
This ability to refer globally is why the export of data from RDBs to the SW should be facilitated, as noted above, by exporting database objects as first-class objects identified with URIs. A number of SW applications have diverged from this vision by not releasing their data onto the Web, but instead archiving them in inaccessible files (at least sometimes because of privacy concerns); Berners-Lee in particular has complained about this tendency [35]. However, this has happend at least sometimes because of privacy concerns.

The next layer up from names is that of markup and data interchange – the realm of XML and RDF. The eXtensible Markup Language (XML [44]) is a metalanguage for markup – in other words, a way of supporting communication and data interchange within communities by defining specialized vocabularies – commonly used in a number of sectors.

XML, like the Hypertext Markup Language (HTML) that underpins the current Web, is descended from the Standard Generalized Markup Language (SGML), an international standard for defining system-independent methods of representing information, and so has no conceptual connection to the SW. The main language for data interchange on the SW, on the other hand, RDF, is specially designed for the task, by assigning specific URIs to the fields in its triples. Figure 3 shows an RDF graph of nodes and arcs made up of several triples – each triple consists of two labeled nodes, from which is pointing a labeled directed arc. The two nodes are the first and third elements of the triple; the arc is the second (it points *from* the first element *to* the third). The use of URIs to refer to the properties as well as the objects is an important step to providing semantics – it enables us to reason about and link to relationships as well as objects.

Figure 3 shows four triples, all 'about' an individual called Eric Miller, identified by 'http://www.w3.org/People/EM/contact#me'. If we look at these triples clockwise from the right, the first represents a connection between Miller, the property of 'having the name ...' (which is referred to by the URI 'http://www.w3.org/2000/10/swap/pim/contact#fullName'), and a character string 'Eric Miller'. The second links Miller, by the property of having a mailbox, to the value of that property, which is his email address given using the common 'mailto:' URI scheme. The third again links Miller with a personal title, the property given as an http URI, and the title as a character string. The fourth triple provides some vocabulary in RDF – it refers to a namespace (an RDF document defining expressive resources which are imported by the RDF graph in Fig. 3) which defines some important relations – and in effect says that Miller (the first object in the triple) is an instance of (a relationship which is the second object in the triple) a person (a class which is the third object in the triple).

Based as it is on triples, RDF is simple yet powerful, exploiting the resources of the common subject/predicate/object structure, and its basis in URIs is very important for the SW. It is a minimalist knowledge representation language for the Web – there are some types of knowledge that cannot be represented in RDF, or only represented with difficulty. For instance, a predicate with more than two arguments has to be represented in a somewhat awkward way as a conjunction of two-argument predicates, while statements about hierarchical class relationships, say, need a further formalism. Furthermore, although the graph structure is quite intuitive, the actual


http://www.w3.org/2000/10/swap/pim/contact#Person
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://www.w3.org/People/EM/contact#me
http://www.w3.org/2000/10/swap/pim/contakt#fullName
Eric Miller
http://www.w3.org/2000/10/swap/pim/contakt#mailbox
mailto:em@w3.org
http://www.w3.org/2000/10/swap/pim/contakt#personalTitle
Dr.

Semantic Web, Figure 3
An RDF graph representing Eric Miller [83]

syntax of RDF is based on XML (it is called RDF/XML) and, although it is well-suited to machine processing, it is not very easy for the human to read (see especially pp. 68–69 in [27]).

The growth in use of RDF has led to the need for special-purpose data stores for holding large quantities of RDF triples (often a set of data will be represented by millions of triples). Such data stores are known as *triplestores*, and need to provide not only storage, but efficient means of reasoning over and retrieving the data that will scale to the large sizes that will be needed. Examples of triplestores include JENA [4], 3store [7,67] and Oracle 11*g* [16].

Greater expressivity is required than is given in RDF, and to that end there another layer upwards which allows the expression of important information about the vocabularies used to express data. As we can see in Fig. 2, the layer here is relatively complex, and contains four boxes, RDFS, Ontology, Rules and Query. These between them provide representation and capabilities that are essential for putting the SW to use.

RDF Schema (RDFS, and sometimes RDF(S) [47]) provides a basic set of tools for producing structured vocabularies that allow different users to agree on particular uses of terms. An extension of RDF, it adds a few modelling primitives with a fixed meaning (such as class, subclass and property relations, and domain and range restriction). It is a basic ontology language that has been adopted fairly widely, and although fairly minimal it can express important constraints on vocabularies.

RDFS is deliberately minimal, and concentrates on expressing subclass and property hierarchies, with various restrictions on these, but the research community, including the Web Ontology Working Group of W3C, identified a number of requirements for greater expressivity for ontologies. As a couple of examples, RDFS allows the stating of subclass relationships, but not, say, that two classes are disjoint; neither does it allow class cardinality restrictions (e. g. a person has *exactly two* parents). As can be seen in Fig. 1, early research efforts into ontology languages led to two leading candidates being developed: DAML (DARPA Agent Markup Language [84]), and OIL (Ontology Interchange Language or Ontology Inference Layer [58]). These two, combined as DAML+OIL [94], became the seed for the W3C ontology language OWL. Unsubstantiated rumor suggests that the fact that it is called 'OWL' and not 'WOL' is an arcane joke: in A. A. Milne's *Winnie-the-Pooh* stories, Owl, who is wise and unusually literate for a forest-dweller, spells his name W-O-L. It is more likely that this is a post hoc rationalization of the naming decision.

The needs of ontology languages are great and potentially problematic. Expressivity is an issue. Even at the level of RDF, its reification mechanism allows the modeler to make statements about statements – an expressive possibility that can lead to logical problems. RDFS has even more powerful modelling primitives, including 'rdfs:Class', the class of all classes. OWL [85] is a strong language for representing concepts and their relations, and its creators needed to wrestle with the inevitable trade-off between expressivity and efficient reasoning support, with two particular constraints demanded by the Semantic Web. First, there is the strong decentralization and lack of enforcement mechanisms on the Web, so that people cannot be forced to use a language they do not want to. Creating a language this powerful, for all purposes, may well have resulted in it not being used at all. Those who need great expressivity might be inclined to use their own favorite non-standard language, while those who want efficient reasoning might prefer to revert to RDFS. The second constraint is that the layered view of the SW (Fig. 2) makes it desirable that OWL should be an extension of RDFS – it should use the RDF interpretation of classes and properties and add primitives to provide richer expressivity. However, this cannot be the basis for OWL, because the addition of powerful reasoning to the expressive power of RDFS (to define such items as the class of all classes) would result in a language very hard to control.

To pick their way between these various pitfalls, the developers of OWL created three separate languages, OWL Full, OWL DL and OWL Lite [85]. OWL Full is the complete language, a full set of OWL primitives, which can be combined with RDF and RDFS in arbitrary ways. This includes the possibility that an OWL Full ontology could augment or alter the meaning of a pre-defined RDF, RDFS or OWL term (for instance, one could put a cardinality constraint on the size of the class of all classes, thereby limiting the number of possible classes that could be constructed). OWL Full is compatible with RDF, so that any legal RDF document is an OWL Full document. The downside of all this expressivity is that the language is undecidable, which rules out the possibility of complete reasoning support.

OWL DL is intended to be open to computational support, and is a sublanguage of OWL Full (so that any legal OWL DL ontology is a legal OWL Full ontology). OWL DL is so named because it is based on *description logic*, a type of knowledge representation language used to describe the knowledge definitive of a particular application domain [29], and is designed to be complete and decidable (in particular, application of OWL's constructors to each other is restricted). This does mean that full compatibility with RDF is lost: although every legal OWL DL document is a legal RDF document, the inverse is not true.

OWL Lite is a very lightweight language to support users requiring a classification hierarchy with simple constraints. Reasoning support should be relatively efficient, and it is intended to provide a straightforward migration route for bringing thesauri and taxonomies to the SW. A legal OWL Lite ontology is also a legal OWL DL ontology, but the cost of ease of reasoning is a lack of expressivity, so for instance enumerated classes, statements of disjointness and assignment of arbitrary cardinality constraints (i. e. restrictions of class cardinality to any number other than 0 and 1) are disallowed.

Even though the relationship between OWL and RDF is complex, its roots in RDF allow OWL to exploit RDF's linking capabilities to allow ontologies to be distributed across systems. When constructing an ontology in OWL, the developer can refer to terms in other ontologies, which then encourages the sharing of terminology across distributed data sources. Sharing ontologies is not always sufficient when it comes to data sharing – an organization may find that nearly all of an imported ontology is adequate, but it needs extra identifiers and descriptions, and in such a case it should be allowed to add them, rather than build a new ontology from scratch [70]. This ability to assemble distributed ontologies is central to the SW vision.

Expressing ontological relations vital, but having achieved a representation of the domain with semantics, one still needs to make inferences. OWL has some inferential support, such as subsumption and classification, but there are several inferential methods that will be required on the SW. Hence, work is currently ongoing on the Rule Interchange Format (RIF), which is intended to allow a variety of rule-based formalisms, ranging from Horn-clause logics, higher order logics and production systems, to be used [41]. Various insights from Artificial Intelligence (AI) have also been adapted for use for the SW for various purposes, including temporal (time-based) logic, causal logic and probabilistic logics [30,75,101,102].

And given the domain description at its desired level of expressivity and a means of making inferences, then the next important function at this level is the ability to query the data. Query language SPARQL works in effect by constructing a graph of RDF-like triples that may contain variables, which is then matched against the RDF graph to be queried; the query is successful if there is a subgraph of the RDF graph which matches the query when RDF terms are substituted for its variables [96].

Sitting on top of these layers are further layers with a unifying logic, proof systems and trust systems. As can be seen from Fig. 1, these upper layers remain topics of exploratory research.

Trust is perhaps key to widespread application of the SW. If information is being drawn from heterogeneous sources, then it is important that users are able to trust such sources if they are to act on the inferences that result. Trust will of course depend on the criticality of the inferences – trust entails risk, and a risk-averse user will naturally trust fewer sources [42,92]. Measuring trust, however, is a complex problem [62]. A key parameter is that of the provenance of data, a statement of the conditions under which data were produced (including statements about the methods of production and the organization that carried them out). Methods are appearing to describe provenance [63], but more needs to be discovered about how information spreads across the Web, and therefore how it can be tracked and understood [39].

Related issues include respect for intellectual property, and the privacy of data subjects. In each case the reasoning abilities of the SW can be of value, and initiatives are currently under way to try to exploit them [93]. Protocols that allow users to express their own privacy preferences, and to enable those who wish to reuse information to reason about those preferences, are being created under the program of research into the Policy Aware Web [108]. Creative Commons is an initiative for representing copyright policies and preferences based on RDF to promote reuse where possible (current standard copyright assumptions are deliberately restrictive with respect to reuse) [2]. Cryptography protocols to protect information and privacy will also play an important role at all levels, as shown in Fig. 2.

Applications

The top layer of Fig. 2 is that of a user interface and applications. This recognizes the fact that if the SW cannot be used easily, and integrated into people's workflows in order to add value to their informational transactions, then it will not attract a large user base, without which the network effects already seen in the development of the World Wide Web will not transpire. Network effects are those positive benefits that increase in certain communication systems faster than its user base expands. In the same way as a telephone system is of limited value to a handful of people and enormous value to a large number, a few people exposing data to the SW is unlikely to make much of a difference, whereas if scalable SW technologies were applied to something like the quantity of data to be found currently in the Deep Web, the gains would be immense.

One not entirely frivolous way of expressing the need for the top layer of the SW is to say that its user base needs to grow quickly, and what is needed is a 'killer app', in other words an application that will meet a felt need and create a perception of the technology as 'essential'. Less ambitiously, the SW's spread depends not only on having an impressive set of formalisms, but also the tools to use the linked data [26].

Bootstrapping

One particular user issue is the importance of bootstrapping content for the SW. Even if RDF began to be published routinely, the amount of legacy content on the Web would dwarf new data for some time, and to make this legacy accessible to SW technology some automation of the process of creating RDF from other formats is required. CS AKTive Space [98], discussed in more detail below, amassed a large quantity of information about the state of computer science research in the United Kingdom through a relatively laborious process of harvesting information from the webpages of computer science departments in British universities without necessarily acting as a source, and using natural language processing and an ontology to interpret the data. The application was very successful, but the researchers were SW researchers, and the process is likely to be too onerous to be repeated on a large scale by non-experts. Assumptions can be made about webpage structure (for example, about regular layouts generated from a database by an individual website), and tools have been developed to exploit them [82].

An important development in this field is GRDDL (Gleaning Resource Descriptions from Dialects of Languages) which became a W3C recommendation in September 2007, and which allows the extraction of RDF from XML and XHTML (a further markup language) documents using transformations expressed in XSLT, an extensible stylesheet language based on XML. It is hoped that such extraction could allow bootstrapping of some of the hoped-for SW network effects, given the amount of XML and XHTML data in the Deep Web [52].

Annotating documents and data with metadata about their content, provenance and other useful dimensions (even including the emotional dimension to content – [100]) is also important for the effort to bring more content into the range of SW technologies [64]. Multimedia are a particular focus for research into annotation [106]. Manual annotation is a great burden for information holders, and a major initial cost for the SW, so methods of automating annotation have been investigated by a number of research teams in order to increase the quantity of annotated data available without excessive expenditure of resources [64,65,107].

In addition, as a large quantity of the Web is actually written in natural language, some have seen a role for natural language processing (NLP), and information extraction (IE), for analyzing this text statistically. So large is the Web's store of written language (two thousand billion words) that it can function as a corpus which dwarfs the most ambitious attempts of dedicated corpus builders in computational linguistics of only a few years ago [79]. And given this, and the need for automating or semi-automating annotation, NLP techniques, augmented by ontologies and training with humans, can be used to extracted machine-readable structured information from plain text [43,48,49,76]. There have also been attempts to build ontologies using NLP techniques, another of the major anticipated bottlenecks for the SW [45]. See also the Subsect. "Commercial and Non-academic Applications" for some examples of SW applications using NLP.

Application Areas

Predicting which particular applications will succeed is unscientific and usually inaccurate. In fact, as it is the rich information contexts that the web of linked data provides that will increase the value of individual pieces of data, one way in which such growth can be encouraged is to focus on small communities with pressing information-processing requirements, and various more-or-less common goals; such communities can be the 'killer apps', or, more accurately, the early adopters of the technology, exactly as the high energy physics discipline played a vital role in the development of the WWW (cf. e. g. [20]). A series of case studies and use cases is maintained at [21].

The most promising of these communities is *e-science*, the data-driven, computationally-intensive pursuit of science in highly distributed computational environments [72]. Large quantities of data are created by analysis and experiments in disciplines such as particle physics, meteorology and the life sciences. Furthermore, in many contexts, different communities of scientists will come together to perform interdisciplinary work, so that data from various fields (e. g. genomics, clinical drug trials and epidemiology) varying not only in vocabulary, but also in the scale of description, need to be integrated. Many scientific disciplines have created large-scale and robust ontologies for this and other purposes. The most well-known of these is the Gene Ontology, a controlled vocabulary to describe gene and gene product attributes in organisms, and related vocabularies developed by Open Biomedical Ontologies. Others include the Protein Ontology, the Cell Cycle Ontology, MeSH (Medical Subject Headings, used to index life science publications), SNOMED (Systematized Nomenclature of Medicine) and AGROVOC (agriculture, forestry, fisheries and food). For more examples and references see [101].

E-government is another important application area, where heterogeneous information of varying quality is deployed widely. Government information varies in provenance, confidentiality and "shelf life" (some information will be good for decades or even centuries, while other information can be out of date within hours), while it can also have been created by various levels of government (national/federal, regional, state, city, parish). Privacy and security are also obviously important factors in this space. Integrating government information in a timely way is clearly an important challenge (see for instance a pilot study for the United Kingdom's Office of Public Sector Information, exploring the use of SW technologies for disseminating, sharing and reusing data held in the public sector [25]).

Academic Applications

Many applications for the SW have been developed with the specific purpose of bringing the SW to maturity. These are often written up in conferences such as the regular World Wide Web Conferences, the International Semantic Web Conferences (ISWC), the European Semantic Web Conferences (ESWC), as well as several one-off conferences and workshops, and can be found in the proceedings (usually online) of any of these. See also the *Journal of Web Semantics* [22].

One initiative of interest here is the Semantic Web Challenge [1], which runs annually alongside the ISWC. This is a good-natured competition to find applications that show SW technology in the best light and which can act as benchmarks for the research community. These applications, therefore, are to some extent an objective list of applications through the years that use semantic technologies to solve real-world problems involving heterogeneous real-world data. The winners of the SW Challenges from its inception, 2003, to the time of writing, 2007, are as follows.

2003: CS AKTive Space (University of Southampton) is an application to explore the UK Computer Science Research domain across multiple dimensions for multiple stakeholders, allowing the tracking of the activities of all agents from funding agencies to individual researchers, using information harvested from the Web, and mediated through an ontology [98].

2004: Flink (Vrije Universiteit Amsterdam) is a 'Who's Who' of the SW which allows the interrogation of information gathered automatically from Web-accessible resources about researchers who have participated in ISWC conferences [86].

2005: CONFOTO (appmosphere web applications, Germany) is a browsing and annotation service for conference photographs [88].

2006: MultimediaN E-Culture Demonstrator (Vrije Universiteit Amsterdam, Centre for Mathematics and

Computer Science, Universiteit van Amsterdam, Digital Heritage Netherlands and Technical University of Eindhoven) searches, navigates and annotates media collections interactively, using digital representations of items from the collections of several well-known museums and art repositories [99].

2007: Revyu.com (Open University) is a reviewing and rating site specifically designed for the SW, allowing reviews of any kind of resource, content or event to be integrated and interlinked with data from other sources (in particular, other reviews, which proliferate on the Web) [69].

A typical SW application will generate a new ontology for its application domain (e. g. art, as with MultimediaN or computer science, as with CS AKTive Space), and use it to interrogate large stores of data, whether legacy data or freshly harvested. This strand of research is tending to confirm the hypothesis that ontologies have an important role in mediating the integration of data from heterogeneous sources.

Commercial and Non-academic Applications

SW applications are generally presented using custom-built interfaces. This suggests a very important area for future research, the development of scalable visualizers capable of navigating the graph of connected information expressed in RDF. As can be seen, the importance of applications and user interfaces was made clear in the layered SW diagram (Fig. 2). However, we shouldn't expect to 'see' the SW in a special browser, in the way that we can see the Web of Documents through browsers such as Internet Explorer, Netscape or Mozilla Firefox. Rather, SW technologies, facilitating the exploration of data, may well work at the back end of websites to improve the user experience. Examples of such sites pointed to by the W3C include Sun's white paper collection site [18], Nokia's developers' discussion forum [11], Oracle's virtual press room [5], and Harper's online magazine [13].

There is an increasing number of applications supporting deeper querying of linked data. The DBpedia [28] is based on the collaborative encyclopedia Wikipedia created by volunteers, and is intended to extract structured information from Wikipedia allowing much more sophisticated querying. Sample queries given on the DPpedia website include a list of people influenced by Friedrich Nietzsche, and the set of images of American guitarists. DBpedia uses RDF, and is also interlinked with other data sources on the Web. When accessed in late 2007, the DBpedia dataset contained 103 million RDF triples. Other examples of linked data applications include the DBLP bibliography of scientific papers [23], and the GeoNames database which represents descriptions of millions of geographical features in RDF [12].

As well as existing organizations using semantic technologies to improve user experience, and applications exploiting linked data, commercial firms are beginning to appear whose business model is based on the possibilities of the SW. Garlik [24] aims to provide individual consumers with more power over their digital data. It reviews what is held about people, harvesting data from the open Web, and represents this in a people-centric structure. Natural Language Processing is used to find occurrences of people's names, sensitive information, and relations to other individuals and organisations (Declaration of interest: Wendy Hall is Chair of the Garlik Advisory Board). Twine [19] aims to facilitate knowledge and information sharing, and to organize that information using various SW technologies (also, like Garlik, using NLP). Twine's developer Nova Spivack coined the term 'knowledge networking' to describe the sharing process, analogous to the Web 2.0 idea of 'social networking'.

Controversies

The SW has been controversial during its history, with several commentators arguing that it is based upon unrealistic expectations, or repeats the mistakes of other initiatives. The arguments against the SW have tended to appear more in the blogosphere rather than the academic world, perhaps because people in the SW world are genuinely enthusiasts while those without confidence in the SW project are doing other things. The pro-SW website GetSemantic supports a wiki page of arguments against the SW, with references and responses [3]. In this section, we will examine three of the most prominent arguments raised against the SW.

The Semantic Web Repeats the Mistakes of "Good Old-Fashioned Artificial Intelligence"

It has been argued that the SW is basically a throwback to the project to program machine intelligence [77] which was jokingly christened by John Haugeland 'GOFAI' (Good Old-Fashioned AI). GOFAI proved impossible: so much of human intelligence is implicit, context-dependent and situated that writing down everything a computer needs to know to produce output that exhibits human-like intelligence is out of the question [68].

One attempt to work around this problem is the Cyc project, set up in 1984, which aims to produce a gigantic ontology that will encode all common-sense knowledge, in order to support human-like reasoning by machines (i. e. GOFAI) [81]. The project has always aroused controversy, but it is fair to say that over two decades later, Cyc is not nearing completion and has not widely perceived as a solution for the GOFAI problem. The implicit nature of common-sense knowledge arguably makes it impossible to write it all down.

Many commentators have argued that the SW is basically a re-creation of the (misconceived) GOFAI idea, that the aim is to create machine intelligence over the Web, to allow machines to reason about Web content in such a way as to exhibit intelligence [77]. This, however, is a misconception, possibly abetted by the strong focus in the 2001 *Scientific American* article on an agent-based vision of the SW [37], although co-author of that article James Hendler has stated very firmly that he believes that the article was radically misinterpreted, and that no-one "can … say we're advocates of the big AI vision when we explicitly make it clear we're pushing for something else" [71]. The *Scientific American* article states that "Traditional knowledge-representation systems typically have been centralized, requiring everyone to share exactly the same definition of common concepts such as 'parent' or 'vehicle'. But central control is stifling, and increasing the size and scope of such a system rapidly becomes unmanageable" [37].

The SW is not GOFAI reheated, but rather an attempt to facilitate sharing of, and context-based machine reasoning over, content (and therefore the provision of machine-readable data on the Web). The aim is not to bring a single ontology, such as Cyc, to bear on all problems (implicitly defining or anticipating all problems and points of view), but to allow data to be interrogated in ways that were not anticipated by their creators. Different ontologies will be appropriate for different purposes; composite ontologies can be assembled from distributed parts [78,85] and it is frequently very basic ontologies (defining simple terms such as 'customer', 'account number' or 'account balance') that add most value to content. In this respect, the situation in the SW simply mirrors offline life where people from different communities and disciplines can and do interact without making any kind of common *global* ontological commitment [36,39,101]. The engineering challenge, as Berners-Lee et al. argue, is to allow independent consistent data systems to be connected locally without requiring global consistency [40].

Yorick Wilks, accepting that the SW is not an attempt to recreate GOFAI, argues that this is both a gain and a loss: a gain because the knowledge representation structures the SW proposes are computationally tractable, as opposed to the various GOFAI formalisms; a loss because DAML+OIL (and presumably by extension OWL) is less sophisticated than those formalisms, and may not have the representational power for the complexity of the world, whether common-sense or scientific [109]. Equally, as both Wilks and Berners-Lee point out, many in the SW world began their research careers in artificial intelligence, as Shadbolt et al. argue that "it will draw on some key insights, tools and techniques derived from 50 years of AI research" [101].

Ontologies

Ontologies, as we have seen, are vital for the SW vision of a Web of Data, but are perceived by many as expensive to develop and hard to maintain. The ideal conceptual apparatus is relative to the task in hand, and different ontologies are appropriate for different tasks. Classifications are also made relative to some background assumptions, and impose those assumptions onto the resulting ontology. To that extent, the expensive development of ontologies reflects the world view of the ontology builders, not necessarily the users. They are top-down and authoritarian, and therefore opposed to the Web ethos of decentralization and open conversation. They are fixed in advance, and so they don't work very well to represent knowledge in dynamic, situated contexts. [95] argues, for instance, that ontologies do not capture the situated processes of scientific research, the social construction of knowledge or the emergence and evolution of understanding over time, and presents an alternative way of representing this knowledge. [104] implicitly endorses this view, showing how there are issues in biology that OWL DL is not well-equipped to handle.

Other papers have made the point that some types of knowledge are more naturally modeled in ontologies than others, and, while not opposing the use of ontologies, warn against too strong a reliance on them for knowledge representation. [61] argues that ontologies cannot be too ambitious, and attempts to reify the context of an ontology (i. e. to provide context-independent accounts of knowledge) will be undermined by knowledge's situated nature. [91] argues that the social context of knowledge requires application builders to be maximally receptive to diverse types of heterogeneous reasoning, which might use knowledge that is hard to capture in hierarchical structures. See also [46] for a series of short essays debating this point.

A related critical point is that the Web as a decentralized, linked information structure must reflect the pragmatic needs of its large, heterogeneous user base which includes very many people who are naive in their understanding of computing issues. The infrastructure has to be usable widely, which argues for simplicity. The rich linking structure of the Web of Documents, combined with statistically-based search engines such as Google, is much more responsive to the needs of unsophisticated users. The SW, in contrast, demands new information representation, markup and publishing practices, and corporations and information owners need to invest in new technologies. Not only that, but current statistical methods will scale up as the number of users and interactions grows, whereas logic-based methods such as those advocated by the SW, on the other hand, scale less well (cf. e. g. [111]).

The dispute has been fueled by the flowering since 2005 or so of the so-called 'Web 2.0' paradigm (of systems, communities and services facilitating collaboration and information-sharing among users). In particular, it has been argued that the meaningful structures that emerge when sufficiently large numbers of users 'tag' content with key words, structures which have been called 'folksonomies', resulting in a structure of connections and classifications emerging without central control, 'really' express the assumptions of the users, and furthermore in such a way as to respect their familiar patterns of communication and workflow. Meanwhile, ontologies 'really' express the needs of the ontology developers and their sponsors [103].

However, folksonomies are much less expressive than ontologies; they are basically variants on keyword searches. A tag 'SF' may refer to science fiction or San Francisco, even if we make the unrealistic assumption of a monoglot English user community. In a multilingual environment such as the Web, further ambiguity is possible – for instance, 'SF' might refer to the Swiss television station Schweizer Fernsehen. Furthermore, the semantics of Web 2.0 are relatively shallow, with few links and very sparse hierarchies.

When a community is large enough and the benefits clear enough to provide incentives to work together, then a large-scale ontology building and maintenance program is justified. It is true that large fixed costs will tend to skew the effort involved towards authorities who may be unrepresentative [91], but Shadbolt et al. argue explicitly that "the ontologies that will furnish the Semantics for the Semantic Web must be developed, managed, and endorsed by committed practice communities. Whether the subject is meteorology or bank transactions, proteins or engine parts, we need concept definitions we can use" [101].

It is of course an undecided question as to whether this community involvement will transpire, but in a recent note, Berners-Lee argues that such conditions will be perhaps more frequently encountered than sceptics believe. On the broad assumption that the size of an ontology-building team increases on the order of the log of the size of the ontology's user community, and that the resources needed to build an ontology increase on the order of the square of community size, the cost per individual of ontology building will diminish rapidly as community size increases. These assumptions are explicitly intended to be indicative rather than realistic [36].

More to the point, not all ontologies need be of great size and expressive depth. It is certainly not the case that the SW requires a single ontology of all discourse on the model of Cyc. Such an ontology, even if possible, would not scale, and in a decentralized structure like the Web its use could not be enforced. Even in complex scientific domains, [74] argues, using a case study from the field of medical informatics, that ontologies should be firmly based on work practices in the domain. In more mundane applications, we should expect a lot of use of small-scale, *shallow* ontologies defining just a few terms that nevertheless are widely applicable [36].

For example, the machine-readable Friend-of-a-Friend (FOAF) ontology is intended to describe people, their activities and their relations to other people. It is not complex, and publishing a FOAF profile is a fairly simple matter for which there are dedicated tools [15]. The resulting network of people has become very large indeed. A survey performed in 2004 discovered over 1.5 million documents using the FOAF ontology [55].

In any case, ontologies and folksonomies serve different purposes. Folksonomies are based on word tags, whereas the basis for ontology reference is via a URI. One of the main aims of ontology definition is to *remove* ambiguity – not globally, for this may well be impossible, but rather within the particular context of the application. Folksonomies will necessarily inherit the ambiguity from the natural language upon which they are based. Nevertheless, a strong possibility that has been considered is to use cheaply-gathered folksonomies as starting points for ontology development, gradually morphing the Web 2.0 structures into something with greater precision and less ambiguity [73,87].

Symbol Grounding

An important aspect of the SW is that URIs must be interpreted consistently. However, terms and symbols are highly variable in their definitions and use through time and space. The SW project will be boosted by processes whereby URIs are given to objects by communities and individuals, endorsed by the user community, who ensure consistency. Responsible URI 'ownership' is critical to the smooth functioning of the SW [101].

But the process of ensuring a fixed and known link between a symbol and its referent, which has been called *symbol grounding*, is at best hard [66], and at worst impossible [110]. Meanings do not stay fixed, but alter, often imperceptibly. They are delineated not only by logical definitions in terms of necessary and sufficient conditions, but also by procedures, technologies and instrumentation, and alter subtly as practice alters.

Any attempt to fix the reference of URIs is a special case of symbol grounding, and is consequently hard to do globally. Attempting to resist the alteration in community practices and norms, and reformulation of meanings of terms, would be doomed. This is understood by leading developers of the SW, who agree that "communities and practice will change norms, conceptualizations, and terminologies in complex and sociologically subtle ways. We shouldn't be surprised or attempt to resist these reformulations" [101]. But there is an important issue, as the same authors concede. "The issue for a Semantic Web built (in a community-driven way) is to know when parts need revision" [101].

Yorick Wilks has argued that Natural Language Processing techniques are essential for grounding the SW, because of the preponderance of text-based content on the Web. NLP is a vital procedural bridge from texts to knowledge representation, usually via automatic information extraction [109]. Berners-Lee has argued in response to Wilks, at a Web Science Workshop in 2005 that the SW was necessarily based on logic and firm definitions (even if those definitions were imperfect, or highly situated and task-relative), not words, use patterns or statistics. Though meanings are not fully stable, they can be stable *enough* relative to individual applications and in particular contexts to allow the SW approach to work [10]. In the case of large-scale, deep ontologies describing sciences, that perhaps will be where the SW is likely to add most value, the Berners-Lee view is reminiscent of that of Hilary Putnam that scientists are 'guardians' of meaning, who determine the 'true' referent of a word like 'water' [97]. But Berners-Lee agrees that ontologies will need to evolve – some quite quickly, and that such meanings cannot be fixed irrevocably; nevertheless, for the purposes of particular applications, this is unlikely to be a problem in practice [101].

Future Directions

The SW is a work in progress, though Shadbolt et al. argue that the need for shared semantics and a Web of Data have increased, and furthermore that the SW is "attainable" [101]. This final section will sketch some of the anticipated directions of future SW work.

Standards

The most obvious future direction is to continue the research as planned. The development of the SW has been conceived as a tide rolling over a beach, covering some areas fully, enveloping other areas more slowly (Fig. 1). As has been noted, the upper layers of the SW, looking at trust, logic and proof, are relatively underdeveloped, and are the focus for exploratory research at the cutting edge. The lower layers of the SW are in place and deployed widely. The middle layers are more or less in place; OWL and SPARQL are complete, while RIF is progressing, and should become a W3C recommendation in the fullness of time.

The Semantic Grid

Grid computing is a type of distributed computing designed to apply computational power from a number of different distributed, complete computers working in parallel, and in cooperation, on a single problem. For some extremely data-heavy problems requiring a lot of computation (particularly in e-science), grid computing is an important time-saving solution. Particular issues in grid computing include the problems of coordinated resource sharing, distributed problem-solving and the creation of 'virtual organisations' to pool data and share outcomes. The SW, of course, is another distributed computing paradigm where data sharing is a key issue – with the SW, a Web of Data, sharing is the whole point. A third distributed paradigm – software agents – is also a relevant factor.

This synergy has led to a research strand to apply semantic technologies to the problems of grid computing, adding meaning via ontologies and RDF metadata annotations to the grid. Information and services for the grid are thereby given well-defined meaning, which enables the interaction between humans and computers to be better coordinated. In particular, all the components, services and resources are adequately described for machine processing. The use of semantics to describe grid resources is known as the *Semantic Grid*, and research is ongoing [17,53,60,72].

The Policy-Aware Web

As is clear in Fig. 2, trusted systems are very important to the development of the SW. There are two reasons for this. First, if someone is reasoning with heterogeneous data harvested from the Web, then they will need to trust the data they have harvested and are using. As noted above, research is ongoing into methods for specifying the provenance of such data [63]. The second reason is that people will not release their data if they thought it would be misused; the importance of data privacy in our digital age is easily underestimated [93]. The *policy-aware Web* is an initiative designed to rectify this problem.

The assumption behind the policy-aware Web is that inflexible and simplistic security systems and access control for the decentralized environment of the Web has hampered its development. Insufficiently sophisticated controls have made people reluctant to share data, particularly with other parties with which they do not have preexisting information-sharing policies. Furthermore, the Web of Documents is rather coarse-grained for detailed security: the security decision to be made is to grant access to an entire website or page, or not, because policy control mechanisms for access at a finer-grained level aren't available. Thus, despite increasing amounts of useful information residing on the Web in a machine-retrieval form, reluctance to share that information remains.

The aim of policy-aware Web technology is to provide for the publication of access policies in a way that allows significant transparency for sharing among partners without requiring pre-agreement. In addition, greater control over information release can be placed in the hands of the information owner, allowing discretionary (rather than mandatory) access control to flourish. Policies would be another kind of metadata attached to information, and those wishing to use that information would be able to reason about them. For instance, one should be able to specify that the information can only be used by the agent gaining access, and that that agent should not pass the information on. Or it may be specified that the information should be deleted after a certain period of time. Or if it is to be used in a certain manner, then data should be anonymized.

Enforcement of these policies is another matter, but at present the research effort is focused on how to express such policies, and on creating theorem provers to reason about them. The result should be a much more fine-grained security picture, with greater transparency and accountability of information use [108].

Web Science

Although since its inception the Web has revolutionized communication, collaboration and education (particularly within science), relatively little is known about the way it develops. There is a growing feeling among researchers across a number of disciplines that a clear research agenda aimed at understanding the current, evolving and potential Web is needed to assure its continued growth. Such researchers want to model the Web, understand the architectural principles that have provided for its growth, and be as sure as possible that it supports the basic social values of trustworthiness, privacy, and respect for social boundaries, and their solution is to chart out a research agenda that targets the Web as a primary focus of attention [39,40].

This agenda has been dubbed *Web Science*, a combination of analysis of the Web and its dynamics, and synthesis of new languages and protocols. The Web is an engineered space created via formally specified languages but, as humans are the creators of Web pages and links between them, their interactions form emergent patterns in the Web at a macroscopic scale. These human interactions are, in turn, governed by social conventions and laws. Web Science is, therefore, inherently interdisciplinary; its goal is to both understand the growth of the Web and to create approaches allowing new powerful and more beneficial patterns to occur.

Such a research area does not yet exist in a coherent form. Within computer science Web-related research has largely focused on information retrieval algorithms and the algorithms for the routing of information through the underlying Internet. Outside of computing, researchers grow ever more dependent on the Web, but there is no concerted agenda for exploring emerging trends on the Web. Nor are those outside computer science fully engaged with the emerging Web research community to focus more specifically on the needs of science and of society as a whole, while preserving the essential invariants of the Web experience: decentralization to avoid social and technical bottlenecks, openness to the reuse of information in unexpected ways, and freedom and equality of information as it passes across the Web.

Despite excitement about the SW, the majority of the world's data is locked in large data stores and is not published as an open web of inter-referring resources. As a result, the reuse of information has been limited. Substantial research challenges arise in changing this situation. We have already discussed the need for policy controls, and for tools to allow scientists to exploit data when it emerges. But on top of that, releasing data is both a technical and a social problem, and understanding how to free data to the SW is a matter of understanding society in relation to the Web (in social, legal and economic terms) and the Web in relation to society. This is the foundation of the emerging Web Science agenda which it is hoped will inform the development of the SW [101]. The recent foundation of the Web Science Research Initiative (WSRI [9]), a joint venture between the Massachusetts Institute of Technology and the University of Southampton, is intended to drive the agenda on, acting as a focus (e. g. advising in particular on curricula to support it).

Bibliography

Primary Literature

1. http://challenge.semanticweb.org/. Accessed Aug 2008
2. http://creativecommons.org/about/. Accessed Dec 2007
3. http://getsemantic.com/wiki/Arguments_against_the_Semantic_Web. Accessed Dec 2007
4. http://jena.sourceforge.net/. Accessed Dec 2007
5. http://pressroom.oracle.com/. Accessed Dec 2007
6. http://protege.stanford.edu/. Accessed Dec 2007
7. http://sourceforge.net/projects/threestore. Accessed Dec 2007
8. http://swoogle.umbc.edu/. Accessed Dec 2007
9. http://webscience.org/. Accessed Dec 2007
10. http://www.cs.umd.edu/~hendler/2005/WebSciReport.pdf. Accessed Dec 2007
11. http://www.forum.nokia.com/. Accessed Dec 2007
12. http://www.geonames.org/. Accessed Dec 2007
13. http://www.harpers.org/. Accessed Dec 2007
14. http://www.informatik.uni-bremen.de/agki/www/swc/index.html. Accessed Dec 2007
15. http://www.ldodds.com/foaf/foaf-a-matic. Accessed Dec 2007
16. http://www.oracle.com/technology/tech/semantic_technologies/index.html. Accessed Dec 2007
17. http://www.semanticgrid.org/. Accessed Dec 2007
18. http://www.sun.com/servers/wp.jsp. Accessed Dec 2007
19. http://www.twine.com/. Accessed Dec 2007
20. http://www.w3.org/2001/sw/SW-FAQ. Accessed Dec 2007
21. http://www.w3.org/2001/sw/sweo/public/UseCases/. Accessed Dec 2007
22. http://www.websemanticsjournal.org/. Accessed Dec 2007
23. http://www4.wiwiss.fu-berlin.de/dblp/. Accessed Dec 2007
24. https://www.garlik.com/index.php. Accessed Dec 2007
25. Alani H, Dupplaw D, Sheridan J, O'Hara K, Darlington J, Shadbolt N, Tullo C (2007) Unlocking the potential of public sector information with Semantic Web technology. In: Proceedings of the 6th international Semantic Web conference, Busan, 2007. http://iswc2007.semanticweb.org/papers/701.pdf. Accessed Dec 2007
26. Alani H, Kalfoglou Y, O'Hara K, Shadbolt N (2005) Towards a killer app for the Semantic Web. In: Gil Y, Motta E, Benjamins VR, Musen MA (eds) The Semantic Web, proceedings of the international Semantic Web conference, Hiroshima, 2005. Springer, Berlin, pp 829–843
27. Antoniou G, van Harmelen F (2004) A Semantic Web primer. MIT Press, Cambridge
28. Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z (2007) DBpedia: a nucleus for a Web of open data. In: Proceedings of the 6th international Semantic Web conference, Busan, South Korea, 2007. http://iswc2007.semanticweb.org/papers/715.pdf. Accessed Dec 2007
29. Baader F, Calvanese D, McGuinness D, Nardi D, Patel-Schneider P (eds) (2003) The description logic handbook: theory, implementation and applications. Cambridge University Press, Cambridge
30. Baclawski K, Niu T (2005) Ontologies for bioinformatics. MIT Press, Cambridge
31. Barrasa J, Corcho O, Gómez-Pérez A (2004) R_2O, an extensible and semantically based database-to-ontology mapping language. In: 2nd Workshop on Semantic Web and Databases (SWDB2004), Toronto, 2004. http://www.cs.man.ac.uk/~ocorcho/documents/SWDB2004_BarrasaEtAl.pdf. Accessed Dec 2007
32. Baumol WJ (2002) The free-market innovation machine: analyzing the growth miracle of capitalism. Princeton University Press, Princeton
33. Berners-Lee T (1994) Plenary at WWW Geneva 94. http://www.w3.org/Talks/WWW94Tim/. Accessed Dec 2007
34. Berners-Lee T (1998) Relational databases on the Semantic Web. http://www.w3.org/DesignIssues/RDB-RDFhtml. Accessed Dec 2007
35. Berners-Lee T (2006/2007) Linked data. http://www.w3.org/DesignIssues/LinkedData.html. Accessed Dec 2007
36. Berners-Lee T (2007) The fractal nature of the Web. http://www.w3.org/DesignIssues/Fractal.html. Accessed Dec 2007
37. Berners-Lee T, Hendler J, Lassila O (2001) The Semantic Web. Scientific American 284(5):34–43. http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21. Accessed Dec 2007
38. Berners-Lee T, Fielding R, Masinter L (2005) Uniform Resource Identifier (URI): generic syntax. http://gbiv.com/protocols/uri/rfc/rfc3986.html. Accessed Dec 2007
39. Berners-Lee T, Hall W, Hendler JA, O'Hara K, Shadbolt N, Weitzner DJ (2006) A framework for Web Science. Found Trends Web Sci 1(1):1–134 http://www.nowpublishers.com/product.aspx?product=WEB&doi=1800000001. Accessed Dec 2007
40. Berners-Lee T, Hall W, Hendler J, Shadbolt N, Weitzner D (2006) Creating a science of the Web. Science 313(5788):769–771
41. Boley H, Kifer M (2007) RIF basic logic dialect. http://www.w3.org/TR/rif-bld/. Accessed Dec 2007
42. Bonatti PA, Duma C, Fuchs N, Nejdl W, Olmedilla D, Peer J, Shahmehri N (2006) Semantic Web policies – a discussion of requirements and research issues. In: Sure Y, Domingue J (eds) The Semantic Web: research and applications, 3rd European Semantic Web Conference 2006 (ESWC-06), Budva. Springer, Berlin
43. Bontcheva K, Tablan V, Maynard D, Cunningham H (2004) Evolving GATE to meet new challenges in language engineering. Nat Lang Eng 10(3/4):349–373
44. Bray T, Paoli J, Sperberg-McQueen CM, Maler E, Yergeau F (2006) Extensible Markup Language (XML) 1.0 (Fourth Edition). http://www.w3.org/TR/xml/. Accessed Dec 2007
45. Brewster C, Ciravegna F, Wilks Y (2002) User-centred ontology learning for knowledge management. In: Andersson B, Bergholtz M, Johannesson P (eds) Proceedings of the 6th international conference on applications of natural language to information systems. Springer, Berlin, pp 203–207
46. Brewster C, O'Hara K (2004) Knowledge representation with ontologies: the present and future. IEEE Intell Syst 19(1): 72–81
47. Brickley D, Guha RV, McBride B (2004) RDF vocabulary description language 1.0: RDF Schema. http://www.w3.org/TR/rdf-schema/. Accessed Dec 2007
48. Ciravegna F, Dingli A, Guthrie L, Wilks Y (2003) Integrating information to bootstrap information extraction from Web sites. In: IJCAI 2003 Workshop on Information Integration on the Web, in conjunction with the 18th International Joint

Conference on Artificial Intelligence (IJCAI 2003), Acapulco, 2003

49. Ciravegna F, Wilks Y (2003) Designing adaptive information extraction for the Semantic Web in Amilcare. In: Handschuh S, Staab S (eds) (2003) Annotation for the Semantic Web. IOS Press, Amsterdam
50. Codd EF (1970) A relational model of data for large shared data banks. Commun ACM 13(6):377–387
51. Connolly D (2003) Semantic Web update: OWL and beyond. http://www.w3.org/2003/Talks/1017-swup/all.htm. Accessed Dec 2007
52. Connolly D (ed) (2007) Gleaning Resource Descriptions from Dialects of Langages (GRDDL). http://www.w3.org/TR/grddl/. Accessed Dec 2007
53. De Roure D, Sure Y (eds) (2006) The Semantic Grid. J Web Semant 4(2):81–123. http://www.websemanticsjournal.org/navigation.html#4. Accessed Dec 2007
54. Ding L, Finin T, Joshi A, Pan R, Cost RS, Peng Y, Reddivari P, Doshi VC, Sachs J (2004) Swoogle: a search and metadata engine for the Semantic Web. In: Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management. ACM Press, pp 652–659. http://ebiquity.umbc.edu/_file_directory_/papers/115.pdf. Accessed Dec 2007
55. Ding L, Zhou L, Finin T, Joshi A (2005) How the Semantic Web is being used: an analysis of FOAF documents. In: Proceedings of the 38th international conference on system sciences. http://ebiquity.umbc.edu/_file_directory_/papers/120.pdf. Accessed Dec 2007
56. Duerst M, Suignard M (2005) Internationalized Resource Identifiers (IRIs). http://www.ietf.org/rfc/rfc3987.txt. Accessed Dec 2007
57. Fensel D (2004) Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce, 2nd edn. Springer, Berlin
58. Fensel D, van Harmelen F, Horrocks I, McGuinness DL, Patel-Schneider PF (2001) OIL: an ontology infrastructure for the Semantic Web. IEEE Intell Syst 16(2):38–45 http://www.cs.man.ac.uk/~horrocks/Publications/download/2001/IEEE-IS01.pdf. Accessed Dec 2007
59. Fielding R, Irvine UC, Gettys J, Mogul JC, Frystyk H, Masinter L, Leach P, Berners-Lee T (1999) Hypertext Transfer Protocol – HTTP/1.1. http://www.w3.org/Protocols/HTTP/1.1/rfc2616.pdf. Accessed Dec 2007
60. Goble C, Kesselman C, Sure Y (eds) (2005) Proceedings of the Dagstuhl seminar on Semantic Grid: the convergence of technologies. http://drops.dagstuhl.de/portals/index.php?semnr=05271. Accessed Dec 2007
61. Goguen JA (2004) Ontology, ontotheology and society. In: International conference on formal ontology in information systems (FOIS 2004). http://charlotte.ucsd.edu/users/goguen/pps/fois04.pdf. Accessed Dec 2007
62. Golbeck J, Hendler J (2004) Accuracy of metrics for inferring trust and reputation in Semantic Web-based social networks. In: Motta E, Shadbolt N, Stutt A, Gibbins N (eds) Engineering Knowledge in the Age of the Semantic Web, Proceedings of 14th International Conference, EKAW 2004, Whittlebury Hall. Springer, Berlin, pp 116–131
63. Groth P, Jiang S, Miles S, Munroe S, Tan V, Tsasakou S, Moreau L (2006) An architecture for provenance systems. http://eprints.ecs.soton.ac.uk/13216/1/provenanceArchitecture10.pdf. Accessed Dec 2007
64. Handschuh S, Staab S (eds) (2003) Annotation for the Semantic Web. IOS Press, Amsterdam
65. Handschuh S, Staab S, Ciravegna F (2002) S-CREAM – Semi-automatic CREAtion of Metadata. In: Gómez-Pérez A, Benjamins VR (eds) Knowledge engineering and knowledge management: ontologies and the Semantic Web, proceedings of 13th international conference, EKAW 2002, Siguënza. Springer, Berlin, pp 358–372
66. Harnad S (1990) The symbol grounding problem. Physica D 42:335–346 http://users.ecs.soton.ac.uk/harnad/Papers/Harnad/harnad90.sgproblem.html. Accessed Dec 2007
67. Harris S, Gibbins N (2003) 3store: efficient bulk RDF storage. In: Proceedings of the 1st International Workshop on Practical and Scalable Systems, Sanibel Island, Florida. http://km.aifb.uni-karlsruhe.de/ws/psss03/proceedings/harris-et-al.pdf. Accessed Dec 2007
68. Haugeland J (1979) Understanding natural language. J Philos 76:619–632
69. Heath T, Motta E (2007) Revyu.com: a reviewing and rating site for the Web of data. In: Proceedings of the 6th international Semantic Web conference 2007, Busan. http://iswc2007.semanticweb.org/papers/889.pdf. Accessed Dec 2007
70. Heflin J (2004) OWL Web Ontology Language use cases and requirements. http://www.w3.org/TR/webont-req/. Accessed Dec 2007
71. Hendler J (2007) Shirkyng my responsibility. http://www.mindswap.org/blog/2007/11/21/shirkyng-my-responsibility. Accessed Dec 2007
72. Hendler J, de Roure D (2004) E-science: the grid and the Semantic Web. IEEE Intell Syst 19(1):65–71
73. Hendler J, Golbeck J (2008) Metcalfe's law, Web 2.0 and the Semantic Web. J Web Semant 6(1):14–20 http://www.websemanticsjournal.org/papers/2007119/MetcalfsLawGolbeckV6I1.pdf. Accessed Dec 2007
74. Hu B, Dasmahapatra S, Dupplaw D, Lewis P, Shadbolt N (2007) Reflections on a medical ontology. Int J Human-comput Stud 65(7):569–582
75. Huang Z, Stuckenschmidt H (2005) Reasoning with multi-version ontologies: a temporal logic approach. In: Proceedings of the 4th International Semantic Web Workshop, http://www.cs.vu.nl/~heiner/public/ISWC05a.pdf. Accessed Dec 2007
76. Iria J, Ciravegna F (2005) Relation extraction for mining the Semantic Web. In: Dagstuhl seminar on machine learning for the Semantic Web. http://tyne.shef.ac.uk/t-rex/pdocs/dagstuhl.pdf. Accessed Dec 2007
77. Jones KS (2004) What's new about the Semantic Web? Some questions. SIGIR forum 38(2) http://www.sigir.org/forum/2004D/sparck_jones_sigirforum_2004d.pdf. Accessed Dec 2007
78. Kalfoglou Y, Schorlemmer M (2003) Ontology mapping: the state of the art. Knowl Eng Rev 18(1):1–31
79. Kilgarrif A, Grefenstette G (2003) Introduction to the special issue on the Web as corpus. Comput Linguist 29(3):333–348 http://www.kilgarriff.co.uk/Publications/2003-KilgGrefenstette-WACIntro.pdf. Accessed Dec 2007
80. Klyne G, Carroll JJ McBride B (2004) Resource Description Framework (RDF): concepts and abstract syntax. http://www.w3.org/TR/rdf-concepts/. Accessed Dec 2007
81. Lenat DB (1995) Cyc: a large-scale investment in knowledge infrastructure. Commun ACM 38(11):33–38

82. Leonard T, Glaser H (2001) Large scale acquisition and maintenance from the Web without source access. In: proceedings of workshop on knowledge markup and semantic annotation, K-CAP2001. http://eprints.ecs.soton.ac.uk/6185/1/Paper.pdf. Accessed Dec 2007
83. Manola F, Miller E, McBride B (2004) RDF primer. http://www.w3.org/TR/rdf-primer/. Accessed Dec 2007
84. McGuinness DL, Fikes R, Stein LA, Hendler J (2003) DAML-ONT: an ontology language for the Semantic Web. In: Fensel D, Hendler J, Lieberman H, Wahlster W (eds) Spinning the Semantic Web: bringing the World Wide Web to its full potential. MIT Press, Cambridge, pp 65–93
85. McGuinness DL, van Harmelen F (2004) OWL Web Ontology Language overview. http://www.w3.org/TR/owl-features/. Accessed Dec 2007
86. Mika P (2005) Flink: Semantic Web technology for the extraction and analysis of social networks. J Web Semant 3(2):211–223 http://www.websemanticsjournal.org/papers/20050719/document7.pdf. Accessed Dec 2007
87. Mika P (2007) Ontologies are us: a unified model of social networks and semantics. J Web Semant 5(1):5–15
88. Nowack B (2005) CONFOTO: A semantic browsing and annotation service for conference photos. In: Gil Y, Motta E, Benjamins VR, Musen MA (eds) The Semantic Web, proceedings of the international Semantic Web conference 2005, Hiroshima. Springer, Berlin, pp 1067–1070
89. Noy NF, McGuinness DL (2001) Ontology development 101: a guide to creating your first ontology. http://smi.stanford.edu/smi-web/reports/SMI-2001-0880.pdf. Accessed Dec 2007
90. Noy NF, Sintek M, Decker S, Crubezy M, Fergerson RW, Musen MA (2001) Creating Semantic Web contents with Protégé-2000. IEEE Intell Syst 16(2):60–71
91. O'Hara K (2004) Ontologies and technologies: knowledge representation or misrepresentation. SIGIR forum 38(2) http://sigir.org/forum/2004D/ohara_sigirforum_2004d.pdf. Accessed Dec 2007
92. O'Hara K, Alani H, Kalfoglou Y, Shadbolt N (2004) Trust strategies for the Semantic Web. In: Workshop on trust, security and reputation on the Semantic Web, 3rd international Semantic Web conference (ISWC 04), Hiroshima, 2004. http://eprints.ecs.soton.ac.uk/10029/. Accessed Dec 2007
93. O'Hara K, Shadbolt N (2008) The spy in the coffee machine: the end of privacy as we know it. Oneworld, Oxford
94. Patel-Schneider P, Horrocks I, van Harmelen F (2002) Reviewing the design of DAML+OIL: an ontology language for the Semantic Web. In: Proceedings of the 18th National Conference on Artificial Intelligence (AAAI02), Edmonton, 2002. http://www.cs.vu.nl/~frankh/postscript/AAAI02.pdf. Accessed Dec 2007
95. Pike W, Gahegan M (2007) Beyond ontologies: toward situated representations of scientific knowledge. Int J Human-comput Stud 65(7):674–688
96. Prud'hommeaux E, Seaborne A (2008) SPARQL query language for RDF. http://www.w3.org/TR/rdf-sparql-query/. Accessed Aug 2008
97. Putnam H (1975) The meaning of 'meaning'. Philosophical papers vol 2: mind, language and reality. Cambridge University Press, Cambridge
98. Schraefel MMC, Shadbolt NR, Gibbins N, Glaser H, Harris S (2004) CS AKTive Space: representing computer science on the Semantic Web. In: Proceedings of WWW 2004, New York, 2004. http://eprints.ecs.soton.ac.uk/9084/. Accessed Dec 2007
99. Schreiber G, Amin A, van Assem M, de Boer V, Hardman L, Hildebrand M, Hollink L, Huang Z, van Kersen J, de Niet M, Omelayenko B, van Ossenbruggen J, Siebes R, Taekema J, Wielemaker J, Wielinga B (2006) MultimediaN e-culture demonstrator. http://www.cs.vu.nl/~guus/papers/Schreiber06a.pdf. Accessed Dec 2007
100. Schröder M, Zovato E, Pirker H, Peter C, Burkhardt F (2007) W3C emotion incubator group report. http://www.w3.org/2005/Incubator/emotion/XGR-emotion/. Accessed Dec 2007
101. Shadbolt N, Hall W, Berners-Lee T (2006) The Semantic Web revisited. IEEE Intell Syst 21(3):96–101
102. Shafer G (1998) 'Causal logic'. In: Proceedings of IJCAI-98. http://www.glennshafer.com/assets/downloads/articles/article62.pdf. Accessed Dec 2007
103. Shirky C (2005) Ontology is overrated: categories, links and tags. http://www.shirky.com/writings/ontology_overrated.html. Accessed Dec 2007
104. Stevens R, Egaña Aranguren M, Wolstencroft K, Sattler U, Drummond N, Horridge M, Rector A (2007) Using OWL to model biological knowledge. Int J Human-comput Stud 65(7):583–594
105. Stojanovic L, Stojanovic N, Volz R (2002) Migrating data-intensive Web sites into the Semantic Web. Symposium on Applied Computing, Madrid
106. Troncy R, van Ossenbruggen J, Pan JZ, Stamou G, Halaschek-Wiener C, Simou N, Tsouvaras V (2007) Image annotation on the Semantic Web. http://www.w3.org/2005/Incubator/mmsem/XGR-image-annotation/. Accessed Dec 2007
107. Vargas-Vera M, Motta E, Domingue J, Lanzoni M, Stutt A, Ciravegna F (2002) MnM: ontology-driven semi-automatic and automatic support for semantic markup. In: Gómez-Pérez A, Benjamins VR (eds) Knowledge engineering and knowledge management: ontologies and the Semantic Web, proceedings of 13th international conference, EKAW 2002, Siguënza, 2002. Springer, Berlin, pp 379–391
108. Weitzner DJ, Hendler J, Berners-Lee T, Connolly D (2005) Creating a policy-aware Web: discretionary, rule-based access for the World Wide Web. In: Ferrari E, Thuraisingham B (eds) Web and information security. Idea Group Inc, Hershey
109. Wilks Y (2008) The Semantic Web as the apotheosis of annotation, but what are its semantics? IEE Intell Syst 23(3):41–49
110. Wittgenstein L (1953) Philosophical investigations. Basil Blackwell, Oxford
111. Zambonini D (2006) The 7 (f)laws of the Semantic Web. http://www.oreillynet.com/xml/blog/2006/06/the_7_flaws_of_the_semantic_we.html [sic]. Accessed Dec 2007

Books and Reviews

Antoniou G, van Harmelen F (2004) A Semantic Web primer. MIT Press, Cambridge

Berners-Lee T (1999) Weaving the Web: the past, present and future of the World Wide Web by its inventor. Texere Publishing, London

Berners-Lee T, Hendler J, Lassila O (2001) The Semantic Web. Scientific American 284(5):34–43. http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21. Accessed Dec 2007

Berners-Lee T, Hall W, Hendler JA, O'Hara K, Shadbolt N, Weitzner DJ (2006) A framework for Web Science. Found Trends Web Sci 1(1):1–134 http://www.nowpublishers.com/product.aspx?product=WEB&doi=1800000001. Accessed Dec 2007

Berners-Lee T, Hall W, Hendler J, Shadbolt N, Weitzner D (2006) Creating a science of the Web. Science 313(5788):769–771

Fensel D (2004) Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce, 2nd edn. Springer, Berlin

Fensel D, Hendler J, Lieberman H, Wahlster W (eds) (2003) Spinning the Semantic Web: bringing the World Wide Web to its full potential. MIT Press, Cambridge

Shadbolt N, Hall W, Berners-Lee T (2006) The Semantic Web revisited. IEEE Intell Syst 21(3):96-101

Semiclassical Spin Transport in Spin-Orbit Coupled Systems

DIMITRIE CULCER[1,2]

[1] Advanced Photon Source, Argonne National Laboratory, Argonne, USA

[2] Northern Illinois University, De Kalb, USA

Article Outline

Abstract

This article discusses spin transport in systems with spin-orbit interactions and how it can be understood in a semiclassical picture. I will first present a semiclassical wave-packet description of spin transport, which explains how the microscopic motion of carriers gives rise to a spin current. Due to spin non-conservation the definition of the spin current has some arbitrariness. In the second part I will briefly review the physics from a density matrix point of view, which makes clear the relationship between spin transport and spin precession and the important role of scattering.

Glossary

Extrinsic effect An effect which has an explicit dependence on the form or strength of the disorder potential.

Intrinsic effect An effect which does not depend explicitly on the form and strength of the disorder potential.

Semiclassical theory A theory in which a particle's position and momentum are considered simultaneously.

Spin-orbit interaction A relativistic interaction between the spin of a particle and its momentum (which is associated with its orbital motion.)

Steady-state spin current A flow of spins induced by an electric field.

Steady-state spin density A net spin density induced by an electric field.

Definition of the Subject

Spin transport refers to the physical movement of spins across a sample and, if spin were a conserved quantity, one could make a straightforward distinction between spin-up and spin-down charge currents. The recent upsurge of interest in spin transport is, however, motivated by systems in which spin is not conserved due to the presence of spin-orbit interactions, which give rise to spin precession. Here, due to non-conservation of spin the spin current is not well defined [1,2,3]. Spin transport in these cases usually does not involve charge transport as the charge currents in the direction of spin flow cancel out. Finally, in certain materials, spin currents are accompanied by steady-state spin densities. The appearance of a spin density is not a transport phenomenon, but it is a steady-state process and is intimately connected to spin transport.

The word *semiclassical* as used in this work refers to theories which consider the position and momentum of a particle simultaneously. Semiclassical pictures are intuitive and useful in descriptions of transport, particularly in inhomogeneous systems and in spatially dependent fields, which typically vary on length scales much larger than atomic size.

In recent years, steady progress has been made towards realization of convenient semiconducting ferromagnets and spin injection into semiconductors from ferromagnetic metals [4,5,6,7] yet spin injection from a ferromagnetic metal into a semiconductor is hampered by the resistivity mismatch between the two [8]. This is one factor, in addition to basic science, motivating the search for an understanding of the way spins are manipulated electrically. The last few years have seen many experimental advances in spin transport, and spin currents have been measured directly [9,10] and indirectly [11,12,13,14,15].

Introduction

Novel physical phenomena that may lead to improved memory devices and advances in quantum information processing are closely related to spin-orbit interactions [16]. Spin-orbit interactions are present in the band structure and in potentials due to impurity distributions. Spin-orbit coupling is in principle always present in impurity potentials and gives rise to skew scattering. Band structure spin-orbit coupling may arise from the inversion asymmetry of the underlying crystal lattice [17] (bulk inversion asymmetry), from the inversion asymmetry of the confining potential in two dimensions [18] (structure inversion asymmetry), and may be present also in inversion symmetric systems [19].

Although many observations in this entry are general, the discussion will focus on non-interacting spin-1/2 electron systems, which are pedagogically easier. The Hamiltonian of these systems typically contains a kinetic energy term and a spin-orbit coupling term, $H_{\boldsymbol{k}} = (\hbar^2 k^2)/(2m^*) + H^{\mathrm{so}}_{\boldsymbol{k}}$, where m^* is the electron effective mass. In spin-1/2 electron systems, band structure spin-orbit coupling can always be represented as a Zeeman-like interaction of the spin with a wave vector-dependent effective magnetic field $\boldsymbol{\Omega}_{\boldsymbol{k}}$, thus $H^{\mathrm{so}}_{\boldsymbol{k}} = (1/2)\boldsymbol{\sigma} \cdot \boldsymbol{\Omega}_{\boldsymbol{k}}$. Common examples of effective fields are the Rashba spin-orbit interaction, [18] which is often dominant in quantum wells with inversion asymmetry, and the Dresselhaus spin-orbit interaction, [17] which is due to bulk inversion asymmetry. The spin operator is given by $s^\sigma = (\hbar/2)\sigma^\sigma$, where σ^σ is a Pauli spin matrix. The spin current operator in these systems will be taken to be $\hat{\jmath}^\sigma_i = (1/2)\{s^\sigma, v^i\}$, where the velocity operator is $v^i = (1/\hbar)\partial H_{\boldsymbol{k}}/\partial k_i$.

An electron spin at wave vector $\mathbf{k}$ precesses about the effective field $\boldsymbol{\Omega}_{\boldsymbol{k}}$ with frequency $\Omega_{\boldsymbol{k}}/\hbar \equiv |\boldsymbol{\Omega}_{\boldsymbol{k}}|/\hbar$ and is scattered to a different wave vector within a characteristic momentum scattering time τ_p. I will assume in this work that $\varepsilon_F \tau_p/\hbar \gg 1$, where ε_F is the Fermi energy, which is equivalent to the assumption that the carrier mean free path is much larger that the de Broglie wavelength. Within this range, the relative magnitude of the spin precession frequency $\Omega_{\boldsymbol{k}}$ and inverse scattering time $1/\tau_p$ define three qualitatively different regimes. In the ballistic (clean) regime no scattering occurs and the temperature tends to absolute zero, so that $\varepsilon_F \tau_p \to \infty$ and $\Omega_{\boldsymbol{k}} \tau_p/\hbar \to \infty$. The weak scattering regime is characterized by fast spin precession and little momentum scattering due to, e. g., a slight increase in temperature, yielding $\varepsilon_F \tau_p/\hbar \gg \Omega_{\boldsymbol{k}} \tau_p/\hbar \gg 1$. In the strong momentum scattering regime $\varepsilon_F \tau_p/\hbar \gg 1 \gg \Omega_{\boldsymbol{k}} \tau_p/\hbar$. I will concentrate on effects originating in the band structure, the observation of which requires the assumption that the materials under study are in the weak momentum scattering regime. Electric fields will be assumed uniform.

The first part of this article will present a semiclassical theory of spin transport, identifying the terms responsible for spin currents in the microscopic dynamics of carriers. Spin non-conservation as a result of spin precession leads to several possible definitions of the spin current, which emerge out of the spin equation of continuity. The second part presents a different point of view, which explains aspects not easily captured in the semiclassical approach. The steady-state density matrix is shown to contain a contribution due to precessing spins and one due to conserved spins. Steady state corrections $\propto \tau_p$ are associated with the *absence* of spin precession and give rise to spin densities in external fields [20,21,22,23,24,25,26,27,28]. Steady state corrections independent of τ_p are associated with spin precession and give rise to spin currents in external fields [1,2,3,9,10,11,12,13,14,15,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46]. Scattering between these two distributions induces significant corrections to steady-state spin currents.

Spin Currents in Electric Fields

Wave-Packet Picture of Spin Transport

This section presents a semiclassical theory of spin transport valid for a general spin-orbit system. The semiclassical method is a suitable approach to the study of transport, because, typically, in the relevant systems the external fields vary smoothly on atomic length scales. All information about the system is taken to be contained in the band structure, thus allowing a description of spin transport which does not make reference to the detailed form of the spin-orbit interaction.

The system under study is regarded as as a collection of carriers, whose semiclassical dynamics in a non-degenerate band i are described by a wave packet [47], with its charge centroid having coordinates $(\mathbf{r}_c, \mathbf{k}_c)$

$$|w_i\rangle = \int \mathrm{d}^3 k \; a(\mathbf{k}, t) \mathrm{e}^{i\mathbf{k}\cdot\hat{\mathbf{r}}} |u_i(\mathbf{r}_c, \mathbf{k}, t)\rangle \; . \qquad (1)$$

In the above, the function $a(\mathbf{k}, t)$ is a narrow distribution sharply peaked at $\mathbf{k}_c$, the phase of which specifies the center of charge position $\mathbf{r}_c$, while $|u_i(\mathbf{r}_c, \mathbf{k}, t)\rangle$ are lattice-periodic Bloch wave functions. The size of the wave packet in momentum space must be considerably smaller than that of the Brillouin zone. In real space, this implies that the wave packet must stretch over many unit cells.

The external electric field drives the center of the wave packet in k-space according to the semiclassical equations of motion

$$\begin{aligned} \hbar\dot{\mathbf{r}}_c &= \frac{\partial \varepsilon_i}{\partial \mathbf{k}_c} - q\mathbf{E} \times \boldsymbol{\Omega}_i \\ \hbar\dot{\mathbf{k}}_c &= q\mathbf{E} \; , \end{aligned} \qquad (2)$$

with q the charge of the carriers, ε_i the band energy, and the Berry curvature

$$\boldsymbol{\Omega}_i = i\left\langle \frac{\partial u_i}{\partial \mathbf{k}} \right| \times \left| \frac{\partial u_i}{\partial \mathbf{k}} \right\rangle . \tag{3}$$

The electric field also gives rise to an adiabatic correction to the wave functions, which mixes the states making up the wave packet. The wave functions $|u_i\rangle$ therefore have the following form:

$$|u_i\rangle = |\phi_i\rangle - \sum_j \frac{\langle \phi_j | i\hbar \frac{\mathrm{d}}{\mathrm{d}t} | \phi_i \rangle}{\varepsilon_i - \varepsilon_j} |\phi_j\rangle \, , \tag{4}$$

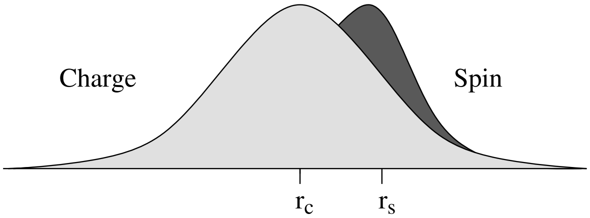
where the ϕ_i are the unperturbed Bloch eigenstates. The $|u_i\rangle$ form a complete set and retain the Bloch periodicity.

The distribution of carriers is described by a function f. When scattering is present, the distribution function satisfies the following equation:

$$\frac{\partial f}{\partial t} + \dot{\mathbf{r}}_c \cdot \frac{\partial f}{\partial \mathbf{r}_c} + \dot{\mathbf{k}}_c \cdot \frac{\partial f}{\partial \mathbf{k}_c} = \left(\frac{\mathrm{d}f}{\mathrm{d}t} \right)_{\mathrm{coll}} , \tag{5}$$

where $(\frac{\mathrm{d}f}{\mathrm{d}t})_{\mathrm{coll}}$ is the usual collision term. In independent bands, in the relaxation time approximation, the collision term takes the form $(f_0 - f)/\tau_p$, with f_0 the equilibrium distribution and τ_p the momentum relaxation time. In the Boltzmann theory, the change in the distribution function with time arises through the drift terms, which are determined from the semiclassical equations of motion, as well as through scattering with other carriers, with localized impurities or with phonons. For transport in a nondegenerate band, it is consistent to ignore interband scattering effects in the weak scattering limit. In this case the relaxation time is a scalar quantity. The effects of interband coherence due to scattering will be explored in the next section.

In order to obtain expressions for macroscopic quantities of interest, such as densities and currents, one needs to carry out a coarse graining by averaging over microscopic fluctuations. In classical dynamics this coarse graining is performed by means of a sampling function, which is smooth and has a significant magnitude only in a finite range [48]. This range is large compared to atomic dimensions, but small compared to the scale of variation of the distribution function. Moreover, it has a rapidly converging Taylor expansion over distances of atomic dimensions, and its form does not need to be specified. This method has a close analog in wavepacket dynamics, where the sampling function is replaced by a δ-function.



Semiclassical Spin Transport in Spin-Orbit Coupled Systems, Figure 1
For a particle of finite extent the charge and spin distributions in real space in general do not coincide. The same is true of the charge and spin distributions in reciprocal space

It is crucial to recognize that, in general, the center of spin and the center of charge are distinct (Fig. 1), since the wave packet samples a range of wave vectors and the spin is usually a function of $\mathbf{k}$. Following the line of thought outlined above, the spin density is defined to be (henceforth $\mathbf{k}_c$ will be abbreviated to $\mathbf{k}$)

$$S^\sigma(\mathbf{R}, t) = \iint \mathrm{d}^3k \mathrm{d}^3r_c f(\mathbf{r}_c, \mathbf{k}, t) \, \langle \delta(\mathbf{R} - \hat{\mathbf{r}}) \hat{s}^\sigma \rangle \, , \tag{6}$$

where the bracket indicates quantum mechanical averaging over the wave packet with charge centroid $(\mathbf{r}_c, \mathbf{k})$. As the δ-function has operator arguments, it will be regarded as a *sampling operator*, whose expectation value yields a spatial average, evaluated at position $\mathbf{r}$. To account for the fact that spin is not conserved (Fig. 2), a new quantity is introduced, which will be referred to as the torque density, defined by

$$\mathcal{T}^\sigma(\mathbf{R}, t) = \iint \mathrm{d}^3k \mathrm{d}^3r_c f(\mathbf{r}_c, \mathbf{k}, t) \, \langle \delta(\mathbf{R} - \hat{\mathbf{r}}) \hat{\tau}^\sigma \rangle \, . \tag{7}$$

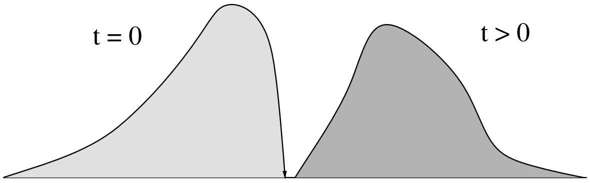
$\hat{\tau}^\sigma$ in the above stands for the rate of change of the spin operator, given by $i/\hbar[\hat{H}, \hat{s}^\sigma]$, and symmetrization of products of non-commuting operators has been assumed. Finally, the microscopic spin current density is defined as:

$$\boldsymbol{\mathcal{J}}^\sigma(\mathbf{R}, t) = \iint \mathrm{d}^3k \mathrm{d}^3r_c f(\mathbf{r}_c, \mathbf{k}, t) \langle \delta(\mathbf{R} - \hat{\mathbf{r}}) \hat{s}^\sigma \hat{\mathbf{v}} \rangle \, . \tag{8}$$

We obtain the following continuity equation for the spin density and current:

$$\frac{\partial S^\sigma}{\partial t} + \nabla \cdot \boldsymbol{\mathcal{J}}^\sigma = \mathcal{T}^\sigma + \mathcal{F}^\sigma \, . \tag{9}$$

The equation of continuity contains a bulk source term, which coincides with the torque density and acts as a mechanism for spin generation. Similar source terms are associated with nonconserved quantities, for example,



Semiclassical Spin Transport in Spin-Orbit Coupled Systems, Figure 2
In the presence of spin-orbit interactions the spin distribution of a particle changes in time. The horizontal axis may represent position or wave vector

in quantum electrodynamics and in Maxwell's equations. The last term in (9) represents the scattering contribution, which will be discussed further below.

Let us discuss the terms in the equation of continuity, beginning with the spin density. The argument of the sampling operator can be expressed as $[\mathbf{r}-\mathbf{r}_c-(\hat{\mathbf{r}}-\mathbf{r}_c)]$, and, as the second term is of atomic dimensions, the sampling operator can be written as a Taylor expansion about $(\hat{\mathbf{r}}-\mathbf{r}_c)$. The density can therefore be re-expressed, in terms of macroscopic quantities, as

$$S^{\sigma}(\mathbf{R},t)=\rho^{s\sigma}(\mathbf{R},t)-\nabla_{\mathbf{R}}\cdot\mathbf{P}^{s\sigma}(\mathbf{R},t)\,, \tag{10}$$

where summation over repeated indices has been assumed. In the above, the monopole density is given by

$$\begin{aligned}\rho^{s\sigma}(\mathbf{R},t)&=\iint \mathrm{d}^3k\mathrm{d}^3r_c f(\mathbf{r}_c,\mathbf{k},t)\langle\hat{s}^{\sigma}\rangle\delta(\mathbf{R}-\mathbf{r}_c)\\&=\int \mathrm{d}^3k f\langle\hat{s}^{\sigma}\rangle|_{\mathbf{r}_c=\mathbf{R}}\,,\end{aligned}\tag{11}$$

where f in the second line, and henceforth, is to be understood as $f(\mathbf{R},\mathbf{k},t)$, and the dipole density is

$$\begin{aligned}\mathbf{P}^{s\sigma}(\mathbf{R},t)&=\iint \mathrm{d}^3k\mathrm{d}^3r_c f\langle(\hat{\mathbf{r}}-\mathbf{r}_c)\hat{s}^{\sigma}\rangle\delta(\mathbf{R}-\mathbf{r}_c)\\&=\int \mathrm{d}^3k f\mathbf{p}^{s\sigma}|_{\mathbf{r}_c=\mathbf{R}}\,.\end{aligned}\tag{12}$$

The average spin of the wave packet has been denoted by $\langle\hat{s}^{\sigma}\rangle$, and the spin-dipole is defined to be $\mathbf{p}^{s\sigma}=\langle(\hat{\mathbf{r}}-\mathbf{r}_c)\hat{s}^{\sigma}\rangle|_{\mathbf{r}_c=\mathbf{R}}$. It will be seen that the first term in the density is the average of a monopole density located at $\mathbf{r}_c$, while the dipole term is the average of a point dipole density located at $\mathbf{r}_c$, and similarly for higher orders. The dipole must be understood as the average of the quantum mechanical dipole operator, as an exact analogy with the electric dipole of classical electrodynamics cannot be made. The density can thus be viewed as a collection of point multipoles, located at the centroid of each wave packet. The microscopic distribution of spin is important at the molecular level, but at the macroscopic level the effect of this molecular distribution is replaced by a sum of multipoles. Since the center of spin is different from the center of charge, in principle all multipoles are present.

Following a similar manipulation and using the Boltzmann equation, the torque density is re-expressed as:

$$\mathcal{T}^{\sigma}(\mathbf{R},t)=\rho^{\tau\sigma}(\mathbf{R},t)-\nabla\cdot\mathbf{P}^{\tau\sigma}(\mathbf{R},t)\tag{13}$$

with the torque monopole density

$$\begin{aligned}\rho^{\tau\sigma}(\mathbf{R},t)&=\iint \mathrm{d}^3k\mathrm{d}^3r_c f(\mathbf{r}_c,\mathbf{k},t)\langle\hat{\tau}^{\sigma}\rangle\delta(\mathbf{R}-\mathbf{r}_c)\\&=\int \mathrm{d}^3k f\langle\hat{\tau}^{\sigma}\rangle|_{\mathbf{r}_c=\mathbf{R}}\,,\end{aligned}\tag{14}$$

and the torque dipole density

$$\begin{aligned}\mathbf{P}^{\tau\sigma}(\mathbf{R},t)&=\iint \mathrm{d}^3k\mathrm{d}^3r_c f(\mathbf{r}_c,\mathbf{k},t)\langle(\hat{\boldsymbol{R}}-\mathbf{r}_c)\hat{\tau}^{\sigma}\rangle|_{\mathbf{r}_c=\mathbf{R}}\\&=\int \mathrm{d}^3k f\mathbf{p}^{\tau\sigma}|_{\mathbf{r}_c=\mathbf{R}}\,.\end{aligned}\tag{15}$$

In analogy with the spin dipole, the torque dipole has been defined as $\mathbf{p}^{\tau\sigma}=\langle(\hat{\mathbf{r}}-\mathbf{r}_c)\hat{\tau}^{\sigma}\rangle|_{\mathbf{r}_c=\mathbf{R}}$. The torque density is therefore also a sum of multipole moments, that is, the moments of a point spin source located at $\mathbf{r}_c$. Even in the case when the center of $\langle\hat{s}^{\sigma}\rangle$ coincides with the center of charge, $\langle\hat{\tau}^{\sigma}\rangle$ may not be centered at $\mathbf{r}_c$, with the result that the higher order terms in the torque density are in general present. The second and higher terms of $\mathcal{T}^{\sigma}$ cancel exactly the analogous terms in the continuity equation which come from the current.

Since only the gradient of the spin current appears in the equation of continuity, in the expansion of the sampling operator we keep the leading term

$$\mathcal{J}^{s}(\mathbf{R},t)=\int \mathrm{d}^3k f\langle\hat{\mathbf{v}}\hat{s}^{\sigma}\rangle|_{\mathbf{r}_c=\mathbf{R}}\,.\tag{16}$$

Keeping terms to first order in $(\hat{\boldsymbol{r}}-\mathbf{r}_c)$, the current can be decomposed into the following:

$$\mathcal{J}^{s\sigma}=\mathbf{C}^{s\sigma}+\mathbf{D}^{s\sigma}-\mathbf{P}^{\tau\sigma}\,.\tag{17}$$

The convective term $\mathbf{C}^{s\sigma}$ represents the spin being transported along with the wave packet

$$\begin{aligned}\mathbf{C}^{s\sigma}(\mathbf{R},t)&=\iint \mathrm{d}^3k\mathrm{d}^3r_c f(\mathbf{r}_c,\mathbf{k},t)\dot{\mathbf{r}}_c\langle\hat{s}^{\sigma}\rangle\delta(\mathbf{R}-\mathbf{r}_c)\\&=\int \mathrm{d}^3k f\mathbf{c}^{s\sigma}|_{\mathbf{r}_c=\mathbf{R}}\,,\end{aligned}\tag{18}$$

while $\mathbf{D}^{s\sigma}$ comes from the rate of change of the spin dipole, which has already been introduced. It has the form:

$$\mathbf{D}^{s\sigma}(\mathbf{R},t) = \int \mathrm{d}^3k f \frac{\mathrm{d}\mathbf{p}^{s\sigma}}{\mathrm{d}t}\Big|_{\mathbf{r}_c=\mathbf{R}} . \tag{19}$$

$\mathbf{P}^{\tau\sigma}$ is the torque dipole introduced above. The corresponding monopole term appears in the source term of the continuity equation, as will emerge below. The presence of the torque dipole here is to be contrasted with the absence of an analogous term in classical electrodynamics. There, an electric dipole arises from the placement of two charges a small distance from each other, but the charge itself is conserved.

Finally, we come to the source term in (9). The first part, composed of the torque density, has already been discussed. The second term, denoted by $\mathcal{F}^\sigma$, becomes, in the relaxation time approximation

$$\mathcal{F}^\sigma = \frac{S_0^\sigma - S^\sigma}{\tau_p} = \frac{1}{\tau_p}\int \mathrm{d}^3k (f_0 - f)\langle \hat{s}^\sigma\rangle , \tag{20}$$

where τ_p is the momentum relaxation time and f_0 the equilibrium distribution, which is usually the Fermi–Dirac distribution function.

Based on the continuity equation alone, there is some flexibility in defining the current and the source. In systems in which spin is conserved, the torque density becomes, to first order in $(\hat{\boldsymbol{r}} - \mathbf{r}_c)$, a pure divergence, which can be incorporated into a redefinition of the spin current. This current, henceforth referred to as the spin transport current, is only due to the convective and spin dipole contributions:

$$\mathbf{J}^{t\sigma}(\mathbf{R},t) = \mathbf{C}^{s\sigma}(\mathbf{R},t) + \mathbf{D}^{s\sigma}(\mathbf{R},t) . \tag{21}$$

With respect to this spin transport current, the continuity equation takes the following form:

$$\frac{\partial S^\sigma}{\partial t} + \nabla\cdot\mathbf{J}^{t\sigma} = \frac{\mathrm{d}}{\mathrm{d}t}\int \mathrm{d}^3k f\langle \hat{s}^\sigma\rangle . \tag{22}$$

In the steady state under a constant electric field, the distribution function is composed of an equilibrium part, independent of the field, and a non-equilibrium part, which is first order in the field. Henceforth, terms in the spin current and source which depend on the equilibrium distribution function will be referred to as intrinsic, whereas the terms depending on the nonequilibrium shift in the distribution will be referred to as extrinsic. For example, the integrand in Eq. (16) can be decomposed into a zero order spin-velocity, $\mathbf{v}\langle\hat{s}^\sigma\rangle$, where $\mathbf{v}$ is the usual group velocity of the band, and a first order correction. Therefore, there will be a contribution to the current from the non-equilibrium part of the distribution and the zero order spin-velocity, which has been discussed extensively in previous work [29,30,31,32,33]. There will also be a contribution from the equilibrium distribution and the first order correction to the spin-velocity, which is referred to as the intrinsic contribution. In the wave packet formalism presented here this effect arises from the change in wave functions induced by the electric field, rather than from the change in distribution functions that is responsible for most conventional transport effects. The intrinsic spin current is calculated from (16) using the equilibrium distribution and the expectation values of the spin and spin dipole operators in a Bloch state perturbed to first order in $\mathbf{E}$.

In its turn, the source in (9) can be decomposed into intrinsic and extrinsic contributions. The present entry considers homogeneous systems, so that all the gradient terms vanish, and the torque density is simply $f\langle\hat{\tau}^\sigma\rangle$. The zeroth order contribution to this term is null, as the Bloch wave functions are eigenstates of the Hamiltonian. Thus, to first order in the electric field, we find that $\langle\hat{\tau}^\sigma\rangle$ is simply given by $(e\mathbf{E}/\hbar)\cdot(\partial\langle\hat{s}^\sigma\rangle/\partial\mathbf{k})$. One is thus justified in replacing f by its equilibrium value f_0, in which case this term is purely intrinsic. The second term in the source, $\mathcal{F}^\sigma$, which depends on the nonequilibrium shift in the distribution function, is entirely extrinsic.

The extrinsic source term takes into account the effect of scattering, and is a term which usually appears in the equation of continuity. The intrinsic source accounts for the effect of all spin-nonconserving terms, and must be present even in a clean system, if the Hamiltonian contains spin-dependent contributions. In general, in addition to the rate of change of spin arising from the spin-dependent terms in the Hamiltonian, scattering processes may alter the orientation of the spin, with the result that any one spin component is not conserved, and the orientation of spins is randomized over a longer time period. For a uniform steady-state system, the current is constant and the intrinsic source term must vanish. However, near the boundary of the system, or at an interface with a different semiconductor with (for example) weaker spin-orbit interactions, the spin current driven by an electric field will vary spatially and $\mathcal{T}$ must reach a non-zero value.

Let us take a closer look at the spin dipole and torque dipole, which are seen to be the main mechanisms responsible for generating the spin current. Because of its narrow distribution in $\mathbf{k}$, the mean spin of the wave packet is $\langle w_i|\hat{s}^\sigma|w_i\rangle = \langle u_i|\hat{s}^\sigma|u_i\rangle$, where it is understood that the wave vector of the Bloch function is set at $\mathbf{k}_c$ and $\hat{s}^\sigma$ is an arbitrary projection of the spin vector operator. The spin dipole of the wave packet, defined relative to the charge center of the wave packet is given, in terms of Bloch functions, by the expression:

$$\mathbf{p}_i^{s\sigma} = \frac{i}{2}\left[\left\langle u_i|\hat{s}^\sigma|\frac{\partial u_i}{\partial \mathbf{k}}\right\rangle - \left\langle \frac{\partial u_i}{\partial \mathbf{k}}|\hat{s}^\sigma|u_i\right\rangle\right] - \left\langle u_i|i\frac{\partial u_i}{\partial \mathbf{k}}\right\rangle \langle u_i|\hat{s}^\sigma|u_i\rangle \ . \quad (23)$$

Interestingly, the spin dipole is independent of the wave packet width. The expression is also invariant under a local gauge transformation, in the sense that if $|u_i\rangle$ is modified by a phase factor $e^{i\alpha(\mathbf{k})}$ the spin dipole is unchanged.

The torque dipole term has a special interpretation in the case of spin transport. The rate of change of spin is equivalent to a torque, and the torque dipole represents the moment exerted by this torque about the center of the wave packet. The semiclassical expression for the torque moment is

$$\mathbf{p}_i^{\tau\sigma} = \frac{i}{2}\left[\left\langle u_i|\dot{\hat{s}}^\sigma|\frac{\partial u_i}{\partial \mathbf{k}}\right\rangle - \left\langle \frac{\partial u_i}{\partial \mathbf{k}}|\dot{\hat{s}}^\sigma|u_i\right\rangle\right] - \left\langle u_i|i\frac{\partial u_i}{\partial \mathbf{k}}\right\rangle \left\langle u_i|\dot{\hat{s}}^\sigma|u_i\right\rangle \ . \quad (24)$$

The torque moment has the same gauge invariance properties as the spin dipole, and like the spin dipole it also does not depend on the wave packet width.

It is important to note that the spin current, $\mathcal{J}^\sigma$, can be simplified to:

$$\mathcal{J}^\sigma(\mathbf{R}, t) = \int \mathrm{d}^3k f \,\mathrm{tr}\langle u_i|\hat{s}^\sigma \hat{\mathbf{v}}|u_i\rangle \ , \quad (25)$$

which is the semiclassical equivalent of the Kubo formula for spin currents.

Density Matrix Picture of Spin Transport

Semiclassical theory provides a straightforward, intuitive picture of the way spin currents arise in the course of carrier dynamics in an electric field. The theory was developed for independent bands. It turns out that interband coherence arising from scattering is crucial in spin transport, and is difficult to treat semiclassically. Although the semiclassical theory can be generalized to multiple bands [49,50], it is more instructive to examine spin transport from a different point of view that is closer in outlook to the philosophy underlying the Kubo formula (with which the semiclassical theory agrees.) This will shed some light on additional issues, such as the relationships between spin currents and spin precession, between spin currents and spin densities, the complex effect of disorder and the vanishing of spin current in certain systems.

A large, uniform system of non-interacting spin-1/2 electrons is represented by a one-particle density operator $\hat{\rho}$. The expectation value of an observable represented by a Hermitian operator $\hat{O}$ is given by $\mathrm{tr}(\hat{\rho}\hat{O})$ and $\hat{\rho}$ satisfies the quantum Liouville equation

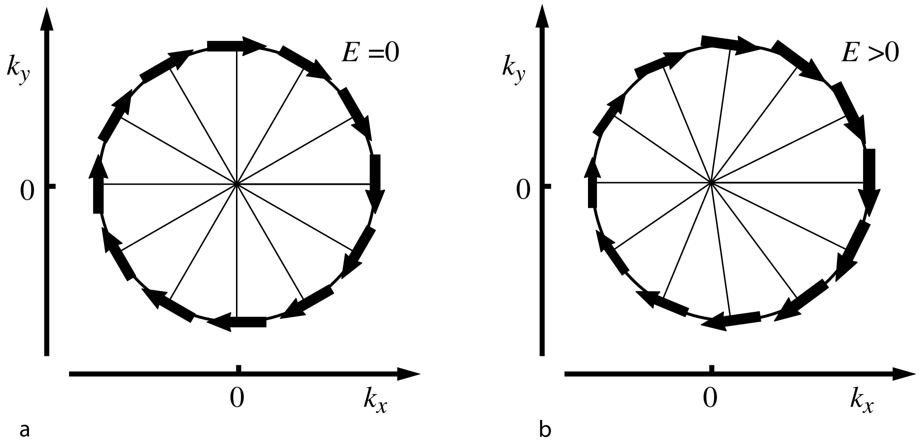
$$\frac{\mathrm{d}\hat{\rho}}{\mathrm{d}t} + \frac{i}{\hbar}[\hat{H} + \hat{U}, \hat{\rho}] = 0\,. \quad (26)$$

The Liouville equation is projected onto a set of time-independent states of definite wave vector $\{|\mathbf{k}\rangle\mathbf{s}\}$, which are not assumed to be eigenstates of the Hamiltonian $\hat{H}$. The matrix elements of $\hat{\rho}$ in this basis will be written as $\rho_{\mathbf{k}\mathbf{k}'} \equiv \rho_{\mathbf{k}\mathbf{k}'}^{ss'} = \langle \mathbf{k}s\hat{\rho}|\mathbf{k}'s'\rangle$. Spin indices are not shown explicitly, and $\rho_{\mathbf{k}\mathbf{k}'}$ is a matrix in spin space, referred to as the density matrix. In this work we require the expectation values of operators which are diagonal in wave vector, and will thus require the part of the density matrix diagonal in wave vector, $\rho_{\mathbf{k}\mathbf{k}} \equiv f_\mathbf{k} = n_\mathbf{k}\mathbb{1} + S_\mathbf{k}$. In the presence of a constant uniform electric field $\mathbf{E}$, $f_\mathbf{k} = f_{0\mathbf{k}} + f_{E\mathbf{k}}$, where the equilibrium density matrix $f_{0\mathbf{k}}$ is given by the Fermi–Dirac distribution, and the correction $f_{E\mathbf{k}}$ is due to the $\mathbf{E}$. We subdivide $f_{0\mathbf{k}} = n_{0\mathbf{k}}\mathbb{1} + S_{0\mathbf{k}}$ and $f_{E\mathbf{k}} = n_{E\mathbf{k}}\mathbb{1} + S_{E\mathbf{k}}$. The spin-dependent part of the nonequilibrium correction to the density matrix $S_{E\mathbf{k}}$ is interpreted as the spin density induced by $\mathbf{E}$. The equations governing the time evolution of $n_{E\mathbf{k}}$ and $S_{E\mathbf{k}}$ is [51]

$$\begin{aligned}\frac{\partial n_{E\mathbf{k}}}{\partial t} + \hat{J}_0(n_{E\mathbf{k}}) &= \frac{e\mathbf{E}}{\hbar}\cdot\frac{\partial n_{0\mathbf{k}}}{\partial \mathbf{k}} \\ \frac{\partial S_{E\mathbf{k}}}{\partial t} + \frac{i}{\hbar}[H_\mathbf{k}, S_{E\mathbf{k}}] + \hat{J}_0(S_{E\mathbf{k}}) &= \frac{e\mathbf{E}}{\hbar}\cdot\frac{\partial S_{0\mathbf{k}}}{\partial \mathbf{k}} - \hat{J}_s(n_{E\mathbf{k}}) \\ &\equiv \Sigma_{E\mathbf{k}}\,,\end{aligned} \quad (27)$$

where the scalar part of the scattering operator $\hat{J}_0$ and its spin-dependent part $\hat{J}_s$ have been defined in [51]. The equation for $n_{E\mathbf{k}}$ has the well-known solution $n_{E\mathbf{k}} = (e\mathbf{E}\tau_p/\hbar)\cdot(\partial n_{0\mathbf{k}}/\partial\mathbf{k})$, in other words, $n_{E\mathbf{k}}$ describes the shift of the Fermi sphere in the presence of the electric field $\mathbf{E}$, with the momentum relaxation time τ_p. It is seen from Eq. (27) that spin-dependent scattering gives rise to a renormalization of the driving term in the equation for $S_{E\mathbf{k}}$. This renormalization has no analog in charge transport.

We need to find the expectation value of the spin current operator $\hat{\mathcal{J}}_i^\sigma$ defined in the introduction. In the systems under study the spin current operator can be written as $\hat{\mathcal{J}}_i^\sigma = \hbar k_i s^\sigma/m^* + (1/4\hbar)\partial\Omega^\sigma/\partial k_i\mathbb{1}$. We need to determine $S_{E\mathbf{k}}$. To this end we remember that an electron spin at wave vector $\mathbf{k}$ precesses about an effective magnetic field $\boldsymbol{\Omega}_\mathbf{k}$. The spin can be resolved into components parallel



Semiclassical Spin Transport in Spin-Orbit Coupled Systems, Figure 3
Effective field $\boldsymbol{\Omega}_k$ at the Fermi energy in the Rashba model [18] a without ($E = 0$) and b with an external electric field ($E > 0$)

and perpendicular to $\boldsymbol{\Omega}_{\mathbf{k}}$. In the course of spin precession the component of the spin parallel to $\boldsymbol{\Omega}_{\mathbf{k}}$ is conserved, while the perpendicular component is continually changing. Corresponding to this decomposition of the spin is an analogous decomposition of the spin distribution $S_{E\mathbf{k}}$ into a part representing conserved spin and a part representing precessing spin, denoted by $S_{E\mathbf{k}\|}$ and $S_{E\mathbf{k}\perp}$ respectively. There is an analogous decomposition of the source on the RHS of Eq. (27) into $\Sigma_{E\mathbf{k}\|}$ and $\Sigma_{E\mathbf{k}\perp}$. This decomposition is carried out by introducing projection operators $P_\|$ and $P_\perp$ as described in [51], giving for $S_{E\mathbf{k}\|}$ and $S_{E\mathbf{k}\perp}$ in the weak momentum scattering limit

$$\frac{\partial S_{E\mathbf{k}\|}}{\partial t} + P_\| \hat{J}_0(S_{E\mathbf{k}}) = \Sigma_{E\mathbf{k}\|} \,, \tag{28a}$$

$$\frac{\partial S_{E\mathbf{k}\perp}}{\partial t} + \frac{i}{\hbar}[H_{\mathbf{k}}, S_{E\mathbf{k}\perp}] = \Sigma_{E\mathbf{k}\perp} - P_\perp \hat{J}_0(S_{E\mathbf{k}}) \,. \tag{28b}$$

Equation (28b) shows that scattering mixes the distributions of conserved and precessing spins. This is so because when one spin at wave vector $\mathbf{k}$ and precessing about $\boldsymbol{\Omega}_{\mathbf{k}}$ is scattered to wave vector $\mathbf{k}'$ and precesses about $\boldsymbol{\Omega}_{\mathbf{k}'}$, its conserved component changes, a process which alters the distributions of conserved and precessing spin. Equations (28a,28b) can be solved straightforwardly if one assumes the impurity potential to be short-ranged, obtaining [51]

$$S_{E\mathbf{k}\|} = \Sigma_{E\mathbf{k}\|}\tau_p + P_\|(1 - \bar{P}_\|)^{-1}\bar{\Sigma}_{E\mathbf{k}\|}\tau_p \,, \tag{29a}$$

$$S_{E\mathbf{k}\perp} = \frac{\boldsymbol{\Omega}_{\mathbf{k}} \times (\boldsymbol{\Sigma}_{E\mathbf{k}\perp}\tau_p + P_\perp \bar{\mathbf{S}}_{E\mathbf{k}\|}) \cdot \boldsymbol{\sigma}\tau_p}{2\hbar(1 + \Omega_{\boldsymbol{k}}^2\tau_p^2/\hbar^2)} - \frac{(\Sigma_{E\mathbf{k}\perp}\tau_p + P_\perp \bar{S}_{E\mathbf{k}\|})}{1 + \Omega_{\mathbf{k}}^2\tau_p^2/\hbar^2} \,. \tag{29b}$$

The correction $S_{E\mathbf{k}\|}$ does not give rise to a spin current. Inspection of Eq. (29a) shows that integrals of the form $\int \mathrm{d}\theta \hat{\jmath}_i^\sigma S_{E\mathbf{k}\|}$ contain an odd number of powers of $\mathbf{k}$ and are therefore zero. It can, however, give rise to a nonequilibrium spin density, since integrals of the form $\int \mathrm{d}\theta \hat{s}^\sigma S_{E\mathbf{k}\|}$ contain an even number of powers of $\mathbf{k}$ and may be nonzero. Similarly $S_{E\mathbf{k}\perp}$ does not lead to a nonequilibrium spin density. The expectation value of the spin operator yields integrals of the form $\int \mathrm{d}\theta \hat{s}^\sigma S^{(0)}_{E\mathbf{k}\perp}$, which involve odd numbers of powers of $\mathbf{k}$ and are therefore zero. This term does, however, give rise to nonzero spin currents, since integrals if the form $\int \mathrm{d}\theta \hat{\jmath}_i^\sigma S_{E\mathbf{k}\perp}$ contain an even numbers of powers of $\mathbf{k}$ and may be nonzero. Therefore, in the absence of spin-orbit coupling in the scattering potential, nonequilibrium spin currents arise from spin precession (as outlined by Sinova et al. [35]), and nonequilibrium spin densities from the absence of spin precession. The dominant contribution to the nonequilibrium spin density in an electric field exists because in the course of spin precession a component of each individual spin is preserved. For an electron with wave vector $\mathbf{k}$, this spin component is parallel to $\boldsymbol{\Omega}_k$. In equilibrium the average of these conserved components is zero. When an electric field is applied the Fermi surface is shifted and the average of the conserved spin components may be nonzero, as illustrated in Fig. 1. This argument explains why the nonequilibrium spin density $\propto \tau_p^{-1}$ and *requires* scattering to balance the drift of the Fermi surface. Although spin densities in electric fields require band structure spin-orbit interactions and therefore spin precession, the dominant contribution arises as a result of the absence of spin precession.

Systems in which $\boldsymbol{\Omega}_{k}$ is linear in $\mathbf{k}$ are special, in that the spin current as defined in this section vanishes [39, 40,41,42,43,44,45,46]. This is because of the renormalization of the driving term on the RHS of Eq. (28b) for $S_{E\mathbf{k}\perp}$, in other words because of scattering from the conserved spin distribution to the precessing spin distribution. In Eq. (29b) it is also clear that if $\Sigma_{E\mathbf{k}\perp}\tau_p + P_{\perp}\bar{S}_{E\mathbf{k}\|}$ vanishes, then all the contributions to $S_{E\mathbf{k}\perp}$ also vanish. Since $\bar{S}_{E\mathbf{k}\|}$ effectively represents a steady-state spin density, we see that the presence of this spin density tends to diminish the spin current. In systems with energy dispersion linear in $\mathbf{k}$ it cancels the spin current completely.

Future Directions

Whereas the community appears to be in agreement that spin currents exist and are measurable, many questions remain unanswered. Theoretically, intrinsic and extrinsic effects (such as due to skew scattering and side jump) have not been studied on the same footing for an arbitrary form of band structure spin-orbit interactions. The relative magnitude of intrinsic and extrinsic spin currents in such a general system remains to be determined. Also, different definitions of the spin current give results that often differ by a sign [1,2]. The relationship between spin current and spin accumulation at the boundary is not clear, again thanks to the non-conservation of spin. It appears that what happens at the boundary is sensitive to the type of boundary conditions assumed. Thus so far as quantitative interpretation of experimental data is concerned, theory has some way to go.

Despite tremendous progress, experiment is still searching for a reliable way to *measure*, as opposed to *detect*, spin currents directly. Practically, the question of what to do with spin once it has been transported/generated remains. The revolutionary electronic device that harnesses spin currents for a practical purpose remains to be made, and the challenge of its design confronts experimentalists and theorists alike.

The research at Argonne National Laboratory was supported by the US Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. DE-AC02-06CH11357.

Bibliography

1. Shi J, Zhang P, Xiao D, Niu Q (2006) Proper definition of spin current in spin-orbit coupled systems. Phys Rev Lett 96:076604
2. Sugimoto N, Onoda S, Murakami S, Nagaosa N (2006) Spin hall effect of a conserved current: Conditions for a nonzero spin hall current. Phys Rev B 73:113305
3. Wang Y, Xia K, Su Z-B, Ma Z (2006) Consistency in formulation of spin current and torque associated with a variance of angular momentum. Phys Rev Lett 96:066601
4. Fiederling R et al (1999) Injection and detection of a spin-polarized current in a light-emitting diode. Nature 402:787
5. Ohno Y et al (1999) Electrical spin injection in a ferromagnetic semiconductor heterostructure. Nature 402:790
6. Jiang X et al (2003) Optical detection of hot-electron spin injection into GaAs from a magnetic tunnel transistor source. Phys Rev Lett 90:256603
7. Jonker BT (2003) Progress toward electrical injection of spin-polarized electrons into semiconductors. Proceedings IEEE 91:727
8. Schmidt G et al (2000) Fundamental obstacle for electrical spin injection from a ferromagnetic metal into a diffusive semiconductor. Phys Rev B 62:R4790
9. Valenzuela SO, Tinkham M (2006) Direct electronic measurement of the spin Hall effect. Nature 442:176
10. Liu B, Shi J, Wang W, Zhao H, Li D, Zhang S-C, Xue Q, Chen D (2006) Experimental observation of the inverse spin hall effect at room temperature. cond-mat/0610150
11. Kato YK, Myers RC, Gossard AC, Awschalom DD (2004) Observation of the spin hall effect in semiconductors. Science 306:1910
12. Wunderlich J, Kaestner B, Sinova J, Jungwirth T (2005) Experimental observation of the Spin-Hall Effect in a two-dimensional Spin-Orbit coupled semiconductor system. Phys Rev Lett 94:047204
13. Sih V, Myers RC, Kato YK, Lau WH, Gossard AC, Awschalom DD (2005) Spatial imaging of the spin Hall effect and current-induced polarization in two-dimensional electron gases. Nature Phys 1:31–35
14. Stern NP, Ghosh S, Xiang G, Zhu M, Samarth N, Awschalom DD (2006) Current-induced polarization and the Spin Hall Effect at room temperature. Phys Rev Lett 97:126603
15. Ganichev SD, Ivchenko EL, Danilov SN, Eroms J, Wegscheider W, Weiss D, Prettl W (2001) Conversion of spin into directed electric current in quantum wells. Phys Rev Lett 86:4358
16. Žutić I, Fabian J, Das Sarma S (2004) Spintronics: Fundamentals and applications. Rev Mod Phys 76:323
17. Dresselhaus G (1955) Spin-Orbit coupling effects in zinc blende structures. Phys Rev 100:580
18. Bychkov YA, Rashba EI (1984) Properties of a 2D electron gas with lifted spectral degeneracy. JETP Lett 39:78
19. Luttinger JM (1956) Quantum theory of cyclotron resonance in semiconductors: General theory. Phys Rev 102:1030
20. Ivchenko EL, Pikus GE (1978) New photogalvanic effect in gyrotropic crystals. JETP Lett 27:604
21. Levitov LS, Nazarov YV, Eliashberg GM (1985) Magnetoelectric effects in conductors with mirror isomer symmetry. Sov Phys JETP 61(1):133
22. Edelstein VM (1990) Spin polarization of conduction electrons induced by electric current in two-dimensional asymmetric electron systems. Solid State Comm 73:233
23. Aronov AG, Lyanda-Geller YB, Pikus GE (1991) Spin polarization of electrons by an electric current. Sov Phys JETP 73:537
24. Magarill LI, Chaplik AV, Entin MV (2001) Spin response of 2D electrons to a lateral electric field. Semiconductors 35(9):1081
25. Vorob'ev LE, Ivchenko EL, Pikus GE, Farbstein II, Shalygin VA, Sturbin AV (1979) Optical activity in tellurium induced by a current. JETP Lett 29:441

26. Kato YK, Myers RC, Gossard AC, Awschalom DD (2004) Current-induced spin polarization in strained semiconductors. Phys Rev Lett 93:176601
27. Ganichev SD, Danilov SN, Schneider P, Bel'kov VV, Golub LE, Wegscheider W, Weiss D, Prettl W (2004) Can an electric current orient spins in quantum wells? cond-mat/0403641
28. Silov AY, Blajnov PA, Wolter JH, Hey R, Ploog KH, Averkiev NS (2004) Current-induced spin polarization at a single heterojunction. Appl Phys Lett 85:5929
29. D'yakonov MI, Perel' VI (1971) Spin orientation of electrons associated with the interband absorption of light in semiconductors. Sov Phys JETP 33:1053–1059
30. Hirsch JE (1999) Spin Hall Effect. Phys Rev Lett 83:1834
31. Zhang S (2000) Spin Hall Effect in the presence of spin diffusion. Phys Rev Lett 85:393
32. Qi Y, Zhang S (2002) Crossover from diffusive to ballistic transport properties in magnetic multilayers. Phys Rev B 65:214407 (ibid 67:052407 (2003))
33. Engel H-A, Halperin BI, Rashba EI (2005) Theory of Spin Hall Conductivity in n-doped GaAs. Phys Rev Lett 95:166605
34. Murakami S, Nagaosa N, Zhang S-C (2003) Dissipationless quantum spin current at room temperature. Science 301:1348
35. Sinova J, Culcer D, Niu Q, Sinitsyn NA, Jungwirth T, MacDonald AH (2004) Universal intrinsic Spin Hall Effect. Phys Rev Lett 92:126603
36. Adagideli I, Bauer GEW (2005) Intrinsic Spin Hall edges. Phys Rev Lett 95:256602
37. Rashba EI (2004) Sum rules for Spin Hall conductivity cancellation. Phys Rev B 70:201309
38. Zhang S, Yang Z (2005) Intrinsic spin and orbital angular momentum hall effect. Phys Rev Lett 94:066602
39. Inoue J, Bauer GEW, Molenkamp LW (2004) Suppression of the persistent spin Hall current by defect scattering. Phys Rev B 70:041303
40. Dimitrova OV (2005) Spin-Hall conductivity in a two-dimensional Rashba electron gas. Phys Rev B 71:245327
41. Schwab P, Raimondi R (2002) Magnetoconductance of a two-dimensional metal in the presence of spin-orbit coupling. Eur Phys J B 25:483
42. Mishchenko EG, Shytov AV, Halperin BI (2004) Spin current and polarization in impure two-dimensional electron systems with spin-orbit coupling. Phys Rev Lett 93:226602
43. Khaetskii A (2006) Nonexistence of intrinsic Spin Currents. Phys Rev Lett 96:056602
44. Shytov AV, Mishchenko EG, Engel H-A, Halperin BI (2006) Small-angle impurity scattering and the spin hall conductivity in two-dimensional semiconductor systems. Phys Rev B 73:075316
45. Liu SY, Lei XL, Horing NJM (2006) Vanishing spin-hall current in a diffusive Rashba two-dimensional electron system: A quantum Boltzmann equation approach. Phys Rev B 73:035323
46. Mal'shukov AG, Chao KA (2005) Spin Hall conductivity of a disordered two-dimensional electron gas with Dresselhaus spin-orbit interaction. Phys Rev B 71:121308
47. Sundaram G, Niu Q (1999) Wave-packet dynamics in slowly perturbed crystals: Gradient corrections and Berry-phase effects. Phys Rev B 59:14915
48. Jackson JD (1999) Classical Electrodynamics, 3rd edn. Wiley, New York, section 6.6, pp 248–258
49. Culcer D, Yao Y, Niu Q (2005) Coherent wave-packet evolution in degenerate bands. Phys Rev B 72:085110
50. Culcer D, Niu Q (2006) Geometrical phase effects in the Wigner distribution of Bloch electrons. Phys Rev B 74:035209
51. Culcer D, Winkler R (2007) Steady states of spin distributions in the presence of spin-orbit interactions. Phys Rev B 76:245322

Shallow Water Waves and Solitary Waves

Willy Hereman
Department of Mathematical and Computer Sciences,
Colorado School of Mines, Golden, USA

Article Outline

Glossary

Deep water A surface wave is said to be in deep water if its wavelength is much shorter than the local water depth.

Internal wave A internal wave travels within the interior of a fluid. The maximum velocity and maximum amplitude occur within the fluid or at an internal boundary (interface). Internal waves depend on the density-stratification of the fluid.

Shallow water A surface wave is said to be in shallow water if its wavelength is much larger than the local water depth.

Shallow water waves Shallow water waves correspond to the flow at the free surface of a body of shallow water under the force of gravity, or to the flow below a horizontal pressure surface in a fluid.

Shallow water wave equations Shallow water wave equations are a set of partial differential equations that describe shallow water waves.

Solitary wave A solitary wave is a localized gravity wave that maintains its coherence and, hence, its visibility through properties of nonlinear hydrodynamics. Solitary waves have finite amplitude and propagate with constant speed and constant shape.

Soliton Solitons are solitary waves that have an elastic scattering property: they retain their shape and speed after colliding with each other.

Surface wave A surface wave travels at the free surface of a fluid. The maximum velocity of the wave and the maximum displacement of fluid particles occur at the free surface of the fluid.

Tsunami A tsunami is a very long ocean wave caused by an underwater earthquake, a submarine volcanic eruption, or by a landslide.

Wave dispersion Wave dispersion in water waves refers to the property that longer waves have lower frequencies and travel faster.

Definition of the Subject

The most familiar water waves are waves at the beach caused by wind or tides, waves created by throwing a stone in a pond, by the wake of a ship, or by raindrops in a river (see Fig. 1). Despite their familiarity, these are all different types of water waves. This article only addresses shallow water waves, where the depth of the water is much smaller than the wavelength of the disturbance of the free surface. Furthermore, the discussion is focused on gravity waves in which buoyancy acts as the restoring force. Little attention will we paid to capillary effects, and capillary waves for which the primary restoring force is surface tension are not covered.

Although the history of shallow water waves [15,20,22] goes back to French and British mathematicians of the eighteenth and early nineteenth century, Stokes [71] is considered one of the pioneers of hydrodynamics (see [21]). He carefully derived the equations for the motion of incompressible, inviscid fluid, subject to a constant vertical gravitational force, where the fluid is bounded below by an impermeable bottom and above by a free surface. Starting from these fundamental equations and by making further simplifying assumptions, various shallow water wave models can be derived. These shallow water models are widely used in oceanography and atmospheric science.

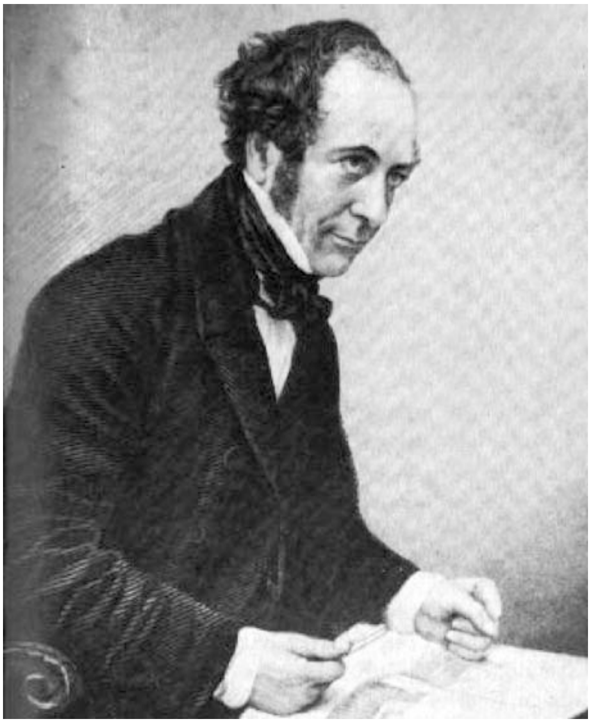
Shallow Water Waves and Solitary Waves, Figure 1
Capillary surface waves from raindrops. Photograph courtesy of E. Scheller and K. Socha

This article discusses shallow water wave equations commonly used in oceanography and atmospheric science. They fall into two major categories: Shallow water wave models with wave dispersion are discussed in Sect. "Completely Integrable Shallow Water Wave Equations" Most of these are completely integrable equations that admit smooth solitary and cnoidal wave solutions for which computational procedures are outlined in Sect. "Computation of Solitary Wave Solutions". Sect. "Shallow Water Wave Equations of Geophysical Fluid Dynamics" covers classical shallow water wave models without dispersion. Such hyperbolic systems can admit shocks. Sect. "Water Wave Experiments and Observations" addresses a few experiments and observations. The article concludes with future directions in Sect. "Future Directions".

Introduction

The initial observation of a solitary wave in shallow water was made by John Scott Russell, shown in Fig. 2. Russell was a Scottish engineer and naval architect who was conducting experiments for the Union Canal Company to design a more efficient canal boat.

In Russell's [63] own words: "I was observing the motion of a boat which was rapidly drawn along a narrow channel by a pair of horses, when the boat suddenly stopped – not so the mass of water in the channel which it had put in motion; it accumulated round the prow of the vessel in a state of violent agitation, then suddenly leaving it behind, rolled forward with great velocity, assuming the form of a large solitary elevation, a rounded, smooth and well-defined heap of water, which continued its course along the channel apparently without change of form or diminution of speed. I followed it on horseback, and overtook it still rolling on at a rate of some eight or nine miles an hour, preserving its original figure some thirty feet long and a foot to a foot and a half in height. Its height gradually diminished, and after a chase of one or two miles I lost it in the windings of the channel. Such, in the month of August 1834, was my first chance interview with that singular and beautiful phenomenon which I have called the Wave of Translation."



Shallow Water Waves and Solitary Waves, Figure 2
John Scott Russell. Source: [27]. Courtesy of John Murray Publishers

Russell built a water tank to replicate the phenomenon and research the properties of the solitary wave he had observed. Details can be found in a biography of John Scott Russell (1808–1882) by Emmerson [27], and in review articles by Bullough [15], Craik [20], and Darrigol [22], who pay tribute to Russell's research of water waves.

In 1895, the Dutch professor Diederik Korteweg and his doctoral student Gustav de Vries [47] derived a partial differential equation (PDE) which models the solitary wave that Russell had observed. Parenthetically, the equation which now bears their name had already appeared in seminal work on water waves published by Boussinesq [13,14] and Rayleigh [59]. The solitary wave was considered a relatively unimportant curiosity in the field of nonlinear waves. That all changed in 1965, when Zabusky and Kruskal realized that the KdV equation arises as the continuum limit of a one dimensional anharmonic lattice used by Fermi, Pasta, and Ulam [29] to investigate "thermalization" – or how energy is distributed among the many possible oscillations in the lattice. Zabusky and Kruskal [78] simulated the collision of solitary waves in a nonlinear crystal lattice and observed that they retain their shapes and speed after collision. Interacting solitary waves merely experience a phase shift, advancing the faster and retarding the slower. In analogy with colliding particles, they coined the word "solitons" to describe these elastically colliding waves. A narrative of the discovery of solitons can be found in [77].

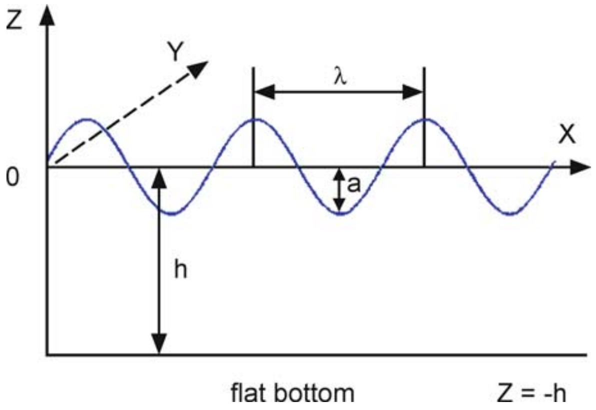
Shallow Water Waves and Solitary Waves, Figure 3
Recreation of a solitary wave on the Scott Russell Aqueduct on the Union Canal. Photograph courtesy of Heriot–Watt University

Since the 1970s, the KdV equation and other equations that admit solitary wave and soliton solutions have been the subject of intense study (see, e. g., [23,30,60]). Indeed, scientists remain intrigued by the physical properties and elegant mathematical theory of the shallow water wave models. In particular, the so-called completely integrable models which can be solved with the Inverse Inverse scattering transform (IST). For details about the IST method the reader is referred to Ablowitz et al. [3], Ablowitz and Segur [2], and Ablowitz and Clarkson [1]. The completely integrable models discussed in the next section are infinite-dimensional Hamiltonian systems, with infinitely many conservation laws and higher-order symmetries, and admit soliton solutions of any order.

As an aside, in 1995, scientists gathered at Heriot–Watt University for a conference and successfully recreated a solitary wave but of smaller dimensions than the one observed by Russell 161 years earlier (see Fig. 3).

Completely Integrable Shallow Water Wave Equations

Starting from Stokes' [71] governing equations for water waves, completely integrable PDEs arise at various levels of approximation in shallow water wave theory. Four length scales play a crucial role in their derivation. As shown in Fig. 4, the wavelength λ of the wave measures the distance between two successive peaks. The amplitude a measures the height of the wave, which is the varying distance between the undisturbed water to the peak of the wave. The



Shallow Water Waves and Solitary Waves, Figure 4
Coordinate frame and periodic wave on the surface of water

depth of the water h is measured from the (flat) bottom of the water up to the quiescent free surface. The fourth length scale is along the Y-axis which is along the crest of the wave and perpendicular to the (X, Z)-plane.

Assuming wave propagation in water of uniform (shallow) depth, i. e. h is constant, and ignoring dissipation, the model equations discussed in this section have a set of common features and limitations which make them mathematically tractable [68]. They describe (i) long waves (or shallow water), i. e. $h \ll \lambda$, (ii) with relatively small amplitude, i. e. $a \ll h$, (iii) traveling in one direction (along the X-axis) or weakly two-dimensional (with a small component in the Y-direction). Furthermore, the small effects must be comparable in size. For example, in the derivation of the KdV and Boussinesq equations one assumes that $\varepsilon = a/h = O(h^2/\lambda^2)$, where ε is a small parameter ($\varepsilon \ll 1$), and O indicates the order of magnitude.

The Korteweg–de Vries Equation

The KdV equation was originally derived to describe shallow water waves of long wavelength and small amplitude. In the derivation, Korteweg and de Vries assumed that all motion is uniform in the Y-direction, along the crest of the wave. In that case, the surface elevation (above the equilibrium level h) of the wave, propagating in the X-direction, is a function only of the horizontal position X (along the canal) and of time T, i. e. $Z = \eta(X, T)$.

In terms of the physical parameters, the KdV equation reads

$$\frac{\partial\eta}{\partial T} + \sqrt{gh}\frac{\partial\eta}{\partial X} + \frac{3}{2}\frac{\sqrt{gh}}{h}\eta\frac{\partial\eta}{\partial X} + \frac{1}{2}h^2\sqrt{gh}\left(\frac{1}{3} - \frac{\mathcal{T}}{\rho g h^2}\right)\frac{\partial^3\eta}{\partial X^3} = 0\,, \quad (1)$$

where h is the uniform water depth, g is the gravitational acceleration (about 9.81m/sec^2 at sea level), ρ is the density, and $\mathcal{T}$ stands for the surface tension. The dimensionless parameter $\mathcal{T}/\rho g h^2$ is called the *Bond number* which measures the relative strength of surface tension and the gravitational force.

Keeping only the first two terms in (1), the speed of the associated linear (long) wave is $c = \sqrt{gh}$. This is indeed the maximum attainable speed of propagation of gravity-induced water waves of infinitesimal amplitude. The speed of propagation of the small-amplitude solitary waves described by (1) is slightly higher. According to Russell's empirical formula the speed equals $\sqrt{g(h+k)}$, where k is the height of the peak of the solitary wave above the surface of undisturbed water. As Bullough [15] has shown, Russell's approximate speed and the true speed of solitary waves only differ by a term of $O(k^2/h^2)$.

The KdV equation can be recast in dimensionless variables as

$$u_t + \alpha u u_x + u_{xxx} = 0\,, \quad (2)$$

where subscripts denote partial derivatives. The parameter α can be scaled to any real number. Commonly used values are $\alpha = \pm 1$ or $\alpha = \pm 6$.

The term u_t describes the time evolution of the wave propagating in one direction. Therefore, (2) is called an *evolution* equation. The nonlinear term $\alpha u u_x$ accounts for steepening of the wave, and the linear dispersive term u_{xxx} describes spreading of the wave. The linear first-order term $\sqrt{gh}\frac{\partial\eta}{\partial X}$ in (1) can be removed by an elementary transformation. Conversely, a linear term in u_x can be added to (2).

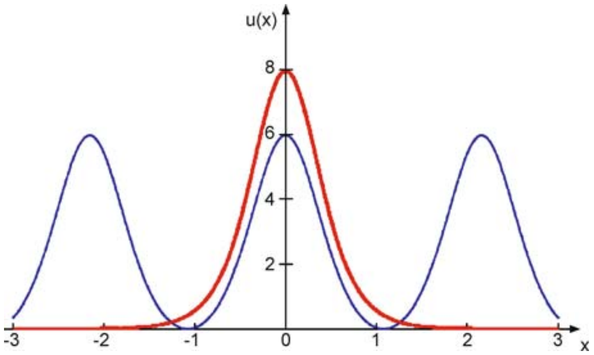
The nonlinear steepening of the water wave can be balanced by dispersion. If so, the result of these counteracting effects is a stable solitary wave with particle-like properties. A solitary wave has a finite amplitude and propagates at constant speed and without change in shape over a fairly long distance. This is in contrast to the concentric group of small-amplitude capillary waves, shown in Fig. 1, which disperse as they propagate.

The closed-form expression of a solitary wave solution is given by

$$u(x,t) = \frac{\omega - 4k^3}{\alpha k} + \frac{12k^2}{\alpha}\operatorname{sech}^2(kx - \omega t + \delta) \quad (3)$$

$$= \frac{\omega + 8k^3}{\alpha k} - \frac{12k^2}{\alpha}\tanh^2(kx - \omega t + \delta)\,, \quad (4)$$

where the wave number k, the angular frequency ω, and δ are arbitrary constants.



Shallow Water Waves and Solitary Waves, Figure 5
Solitary wave (red) and periodic cnoidal (blue) wave profiles

Requiring that $\lim_{x\to\pm\infty} u(x,t) = 0$ for all time leads to $\omega = 4k^3$. Then (3) and (4) reduce to

$$\begin{aligned} u(x,t) &= \frac{12k^2}{\alpha}\operatorname{sech}^2(kx - 4k^3t + \delta) \\ &= \frac{12k^2}{\alpha}[1 - \tanh^2(kx - 4k^3t + \delta)] \,. \end{aligned} \tag{5}$$

The position of the hump-type wave at $t = 0$ is depicted in Fig. 5 for $\alpha = 6,\ k = 2$, and $\delta = 0$. As time changes, the solitary wave with amplitude $2k^2 = 8$ travels to the right at speed $v = \omega/k = 4k^2 = 16$. The speed is exactly twice the peak amplitude. So, the higher the wave the faster it travels, but it does so without change in shape. The reciprocal of the wavenumber k is a measure of the width of the sech-square pulse.

As shown by Korteweg and de Vries [47], Eq. (2) also has a simple periodic solution,

$$u(x,t) = \frac{\omega - 4k^3(2m-1)}{\alpha k} + \frac{12k^2m}{\alpha}\operatorname{cn}^2(kx - \omega t + \delta; m), \tag{6}$$

which they called the *cnoidal wave* solution since they involve the Jacobi elliptic cosine function, cn, with modulus $m\,(0 < m < 1)$. The wavenumber k gives the characteristic width of each oscillation in the "cnoid."

Three cycles of the cnoidal wave are depicted in Fig. 5 at $t = 0$. The graph corresponds to $\alpha = 6, k = 2, m = 3/4, \omega = 16$, and $\delta = 0$. Using the property $\lim_{m\to 1} \operatorname{cn}(\xi; m) = \operatorname{sech}(\xi)$, one readily verifies that (6) reduces to (3) as m tends to 1. Pictorially, the individual oscillations then stretch infinitely far apart leaving a single-pulse solitary wave.

The celebrated KdV equation appears in all books and reviews about soliton theory. In addition, the equation has been featured in, e. g., Miura [52] and Miles [51].

Regularized Long-Wave Equations

A couple of alternatives to the KdV equation have been proposed. A first alternative,

$$u_t + u_x + \alpha u u_x - u_{xxt} = 0 \,, \tag{7}$$

was proposed by Benjamin, Bona, and Mahony [7]. Hence, (7) is referred to as the BBM or regularized long-wave (RLW) equation.

Equation (7), which has a solitary wave solution,

$$u(x,t) = \frac{\omega - k - 4k^2\omega}{\alpha k} + \frac{12k\omega}{\alpha}\operatorname{sech}^2(kx - \omega t + \delta), \tag{8}$$

was also derived by Peregrine [57] to describe the behavior of an undular bore (in water), which comprises a smooth wavefront followed by a train of solitary waves. An undular bore can be interpreted as the dispersive analog of a shock wave in classical non-dispersive, dissipative hydrodynamics [26].

The linear dispersion relation for the KdV equation, $\omega = k(1 - k^2)$, can be obtained by substituting $u(x,t) = \exp[i(kx - \omega t + \delta)]$ into $u_t + u_x + u_{xxx} = 0$. The linear phase velocity, $v_p = \omega/k = 1 - k^2$, becomes negative for $|k| > 1$, thereby contradicting the assumption of unidirectional propagation. Furthermore, the group velocity $v_g = d\omega/dk = 1 - 3k^2$ has no lower bound which implies that there is no limit to the rate at which shorter ripples propagate in the negative x-direction.

The BBM equation, where $\omega = k/(1 + k^2), v_p = 1/(1 + k^2)$, and $v_g = (1 - k^2)/(1 + k^2)^2$, was proposed to get around these objections and to address issues related to proving the existence of solutions of the KdV equation. The dispersion relation of (7) has more desirable properties for high wave numbers, but the group velocity becomes negative for $|k| > 1$. In addition, the KdV and BBM equations are first order in time making it impossible to specify both u and u_t as initial data.

To circumvent these limitations, a second alternative,

$$u_t + u_x + \alpha u u_x + u_{xtt} = 0 \,, \tag{9}$$

was proposed by Joseph and Egri [45] and Jeffrey [43]. It is called the time regularized long-wave (TRLW) equation and its solitary wave solution is given by

$$u(x,t) = \frac{\omega - k - 4k\omega^2}{\alpha k} + \frac{12\omega^2}{\alpha}\operatorname{sech}^2(kx - \omega t + \delta). \tag{10}$$

The TRLW equation shares many of the properties of both the KdV and BBM equations, at the cost of a more complicated dispersion relation, $\omega = (-1 \pm \sqrt{1 + 4k^2})/2k$ with two branches. Only one of these branches is valid because the derivation of the TRLW equation shows that (9) is unidirectional, despite the two time derivatives in u_{xtt}.

Bona and Chen [8] have shown that the initial value problem for the TRLW equation is well-posed, and that for small-amplitude, long waves, solutions of (9) agree with solutions of either (2) or (7). As a matter of fact, all three equations agree to the neglected order of approximation over a long time scale, provided the initial data is properly imposed (see also [9]).

Fine-tuning the dispersion relation of the KdV equation comes at a cost. In contrast to (2), the RLW and TRLW equations are no longer completely integrable. Perhaps that is why these equations never became as popular as the KdV equation.

The Boussinesq Equation

The classical Boussinesq equation,

$$\eta_{TT}-c^2\eta_{XX}-\frac{3c^2}{h}\left(\eta_X^2+\eta\eta_{XX}\right)-\frac{c^2h^2}{3}\eta_{XXXX}=0, \quad (11)$$

was derived by Boussinesq [12] to describe gravity-induced surface waves as they propagate at constant (linear) speed $c=\sqrt{gh}$ in a canal of uniform depth h.

In contrast to the KdV equation, (11) has a second-order time-derivative term. Ignoring all but the first two terms in (11), one obtains the linear *wave* equation, $\eta_{TT}-c^2\eta_{XX}=0$, which describes both left-running and right-running waves. However, (11) is *not* bi-directional because in the derivation Boussinesq used the constraint $\eta_T=-c\eta_X$, which limits (11) to waves traveling to the right.

In dimensionless form, the Boussinesq equation reads

$$u_{tt}-c^2u_{xx}-\alpha u_x^2-\alpha uu_{xx}-\beta u_{xxxx}=0\,. \quad (12)$$

The values of the parameters $c, \alpha>0$, and β do not matter, but the sign of β matters. Typically, one sets $c=1, \alpha=3$, and $\beta=\pm1$.

A simple solitary wave solution of (12) is given by

$$u(x,t)=\frac{\omega^2-c^2k^2-4\beta k^4}{\alpha k^2}+\frac{12\beta k^2}{\alpha}\operatorname{sech}^2(kx-\omega t+\delta)\,. \quad (13)$$

The equation with $\beta=1$ is a scaled version of (11) and thus most relevant to shallow water wave theory. Mathematically, (12) with $\beta=1$ is ill-posed, (even without the nonlinear terms), which means that the initial value problem cannot be solved for arbitrary data. This shortcoming does not happen for (12) with $\beta=-1$, which is therefore nicknamed the "good" Boussinesq equation [50]. Nonetheless, the classical and good Boussinesq equations are completely integrable.

The "improved" or "regularized" Boussinesq equation (see, e. g., [10]) has $\beta=1$ but u_{xxtt} instead of u_{xxxx}, which improves the properties of the dispersion relation. Like (12), the regularized version describes uni-directional waves. The regularized Boussinesq equation and other alternative equations listed in the literature (see, e. g., Madsen and Schäffer [49]) are not completely integrable.

Bona et al. [10,11] analyzed a family of Boussinesq systems of the form

$$\begin{aligned} w_t+u_x+(uw)_x+\alpha u_{xxx}-\beta w_{xxt}&=0\,,\\ u_t+w_x+uu_x+\gamma w_{xxx}-\delta u_{xxt}&=0\,, \end{aligned} \quad (14)$$

which follow from the Euler equations as first-order approximations in the parameters $\varepsilon_1=a/h\ll1$, $\varepsilon_2=h^2/\lambda^2\ll1$, where the Stokes number, $S=\varepsilon_1/\varepsilon_2=a\lambda^2/h^3\approx1$.

In (14) $w(x,t)$ is the non-dimensional deviation of the water surface from its undisturbed position; $u(x,t)$ is the non-dimensional horizontal velocity field at a height θh (with $0\le\theta\le1$) above the flat bottom of the water. The constant parameters α through δ in (14) satisfy the following consistency conditions: $\alpha+\beta=\frac{1}{2}(\theta^2-\frac{1}{3})$ and $\gamma+\delta=\frac{1}{2}(1-\theta^2)\ge0$. Solitary wave solutions of various special cases of (14) have been computed by Chen [18].

Boussinesq systems arise when modeling the propagation of long-crested waves on large bodies of water (such as large lakes or the ocean). The Boussinesq family (14) encompasses many systems that appeared in the literature. Special cases and properties of well-posedness of (14) are addressed by Bona et al. [10,11].

1D Shallow Water Wave Equation

The so-called one-dimensional (1D) shallow water wave equation,

$$v_{xxt}+\alpha vv_t-v_t-v_x+\beta v_x\int_{\infty}^{x}v_t(y,t)dy=0\,, \quad (15)$$

can be derived from the classical shallow water wave theory (see Sect. "Shallow Water Wave Equations of Geophysical Fluid Dynamics") in the Boussinesq approximation. In that approximation one assumes that vertical variations of the static density, ρ_0, are neglected, except the buoyancy term proportional to $d\rho_0/dz$, which is, in fact, responsible for the existence of solitary waves. The integral term in (15) can be removed by introducing the potential u. Indeed, setting $v=u_x$, Eq. (15) can be written as

$$u_{xxxt}+\alpha u_xu_{xt}-u_{xt}-u_{xx}+\beta u_{xx}u_t=0\,. \quad (16)$$

The equation is completely integrable and can be solved with the IST if and only if either $\alpha = \beta$ [41] or $\alpha = 2\beta$ [3]. When $\alpha = \beta$, Eq. (16) can be integrated with respect to x and thus replaced by

$$u_{xxt} + \alpha u_x u_t - u_t - u_x = 0 \,. \tag{17}$$

Closed-form solutions of (15), and in particular of (17), have been computed by Clarkson and Mansfield [19].

The Camassa–Holm Equation

The CH equation, named after Camassa and Holm [16,17],

$$\begin{aligned} u_t + 2\kappa u_x + 3uu_x - \alpha^2 u_{xxt} + \gamma u_{xxx} \\ - 2\alpha^2 u_x u_{xx} - \alpha^2 u u_{xxx} = 0 \,, \end{aligned} \tag{18}$$

also models waves in shallow water. In (18), u is the fluid velocity in the x-direction or, equivalently, the height of the water's free surface above a flat bottom, and κ, γ and α are constants. Retaining only the first four terms in (18) gives the BBM Eq. (7). Setting $\alpha = 0$ reduces (18) to the KdV equation.

The CH equation admits solitary wave solutions, but in contrast to the hump-type solutions of the KdV and Boussinesq equations, they are implicit in nature (see, e. g., [44]). In the limit $\kappa \to 0$, Eq. (18) with $\gamma = 0, \alpha = 1$ has a cusp-type solution of the form $u(x,t) = c\exp(-|x - ct - x_0|)$. The solution is called a peakon since it has a peak (or corner) where the first derivatives are discontinuous. The solution travels at speed $c > 0$ which equals the height of the peakon.

The Kadomtsev–Petviashvili Equation

In their 1970 study [46] of the stability of line solitons, Kadomtsev and Petviashvili (KP) derived a 2D-generalization of the KdV equation which now bears their name. In dimensionless variables, the KP equation is

$$(u_t + \alpha u u_x + u_{xxx})_x + \sigma^2 u_{yy} = 0 \,, \tag{19}$$

where y is the transverse direction. In the derivation of the KP equation, one assumes that the scale of variation in the y-direction (along the crest of the wave as shown in Fig. 4) is much longer than the wavelength along the x-direction.

The solitary wave and periodic (cnoidal) solutions of (19) are, respectively, given by

$$u(x,t) = \frac{k\omega - 4k^4 + \sigma^2 l^2}{\alpha k^2} + \frac{12k^2}{\alpha}\operatorname{sech}^2(kx + ly - \omega t + \delta), \tag{20}$$

Shallow Water Waves and Solitary Waves, Figure 6
Periodic plane waves in shallow water, off the coast of Lima, Peru. Photograph courtesy of A. Segur

and

$$\begin{aligned} u(x,t) = {} & \frac{k\omega - 4k^4(2m-1) - \sigma^2 l^2}{\alpha k^2} \\ & + \frac{12k^2 m}{\alpha}\operatorname{cn}^2(kx + ly - \omega t + \delta; m) \,. \end{aligned} \tag{21}$$

As shown in Fig. 6, near a flat beach the periodic waves appear as long, quasilinear stripes with a cn-squared cross section. Such waves are typically generated by wind and tides.

The equation with $\sigma^2 = -1$ is referred to as KP1, whereas (19) with $\sigma^2 = 1$ is called KP2, which describes shallow water waves [67]. Both KP1 and KP2 are completely integrable equations but their solution structures are fundamentally different (see, e. g., [65], pp. 489–490).

Shallow Water Wave Equations of Geophysical Fluid Dynamics

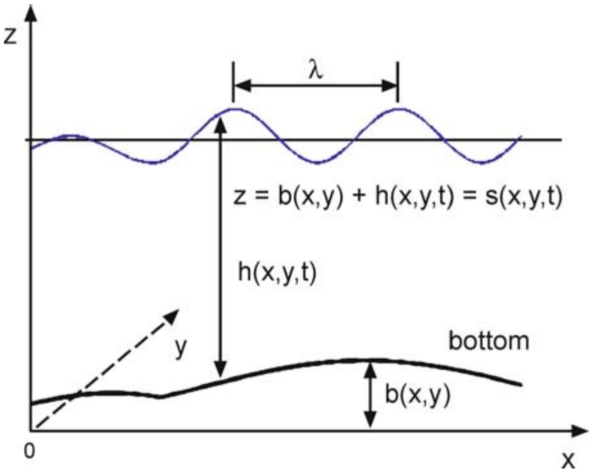
The shallow water equations used in geophysical fluid dynamics are based on the assumption $D/L \ll 1$, where D and L are characteristic values for the vertical and horizontal length scales of motion. For example, D could be the average depth of a layer of fluid (or the entire fluid) and L could be the wavelength of the wave.

The geophysical fluid dynamics community (see, e. g., [55,72,73]) uses the following 2D shallow water equations,

$$u_t + uu_x + vu_y + gh_x - 2\Omega v = -gb_x \,, \tag{22}$$

$$v_t + uv_x + vv_y + gh_y + 2\Omega u = -gb_y, \tag{23}$$

$$h_t + (hu)_x + (hv)_y = 0, \tag{24}$$



Shallow Water Waves and Solitary Waves, Figure 7
Setup for the geophysical shallow water wave model

to describe water flows with a free surface under the influence of gravity (with gravitational acceleration g) and the Coriolis force due to the earth's rotation (with angular velocity Ω.) As usual, $\mathbf{u} = (u, v)$ denotes the horizontal velocity of the fluid and $h(x, y, t)$ is its depth. As shown in Fig. 7, $h(x, y, t)$ is the distance between the free surface $z = s(x, y, t)$ and the variable bottom $b(x, y)$. Hence, $s(x, y, t) = b(x, y) + h(x, y, t)$. Eqs. (22) and (23) express the horizontal momentum-balance; (24) expresses conservation of mass. Note that the vertical component of the fluid velocity has been eliminated from the dynamics and that the number of independent variables has been reduced by one. Indeed, z no longer explicitly appears in (22)–(24), where u, v, and h only depend on x, y, and t.

A shortcoming of the model is that it does not take into account the density stratification which is present in the atmosphere (as well as in the ocean). Nonetheless, (22)–(24) are commonly used by atmospheric scientists to model flow of air at low speed.

More sophisticated models treat the ocean or atmosphere as a stack of layers with variable thickness. Within each layer, the density is either assumed to be uniform or may vary horizontally due to temperature gradients. For example, Lavoie's rotating shallow water wave equations (see [24]),

$$\boldsymbol{u}_t + (\boldsymbol{u} \cdot \nabla)\boldsymbol{u} + 2\boldsymbol{\Omega} \times \boldsymbol{u} = -\nabla(h\theta) + \tfrac{1}{2} h \nabla\theta \,, \qquad (25)$$

$$h_t + \nabla \cdot (h\boldsymbol{u}) = 0 \,, \qquad (26)$$

$$\theta_t + \boldsymbol{u} \cdot (\nabla\theta) = 0 \,, \qquad (27)$$

consider only one active layer with layer depth $h(x, y, t)$, but take into account the forcing due to a horizontally varying potential temperature field $\theta(x, y, t)$. Vector $\boldsymbol{u} = u(x, y, t)\boldsymbol{i} + v(x, y, t)\boldsymbol{j}$ denotes the fluid velocity and $\boldsymbol{\Omega} = \Omega\boldsymbol{k}$ is the angular velocity vector of the Earth's rotation. $\nabla = \frac{\partial}{\partial x}\boldsymbol{i} + \frac{\partial}{\partial y}\boldsymbol{j}$ is the gradient operator, and $\boldsymbol{i}$, $\boldsymbol{j}$, and $\boldsymbol{k}$ are unit vectors along the x, y, and z-axes.

Lavoie's equations are part of a family of multi-layer models proposed by Ripa [61] to study, for example, the effects of solar heating, fresh water fluxes, and wind stresses on the upper equatorial ocean. A study of the validity of various layered models has been presented by Ripa [62]. The more sophisticated the models become the harder they become to treat with analytic methods so one has to apply numerical methods. Numerical aspects of various shallow water models in atmospheric research and beyond are discussed in e. g. [48,75,76].

Computation of Solitary Wave Solutions

As shown in Sect. "Completely Integrable Shallow Water Wave Equations", solitary wave solutions of the KdV and Boussinesq equations (and like PDEs), can be expressed as polynomials of the hyperbolic secant (sech) or tangent (tanh) functions, whereas their simplest period solutions involve the Jacobi elliptic cosine (cn) function.

There are several methods to compute exact, analytic expressions for solitary and periodic wave solutions of nonlinear PDEs. Two straightforward methods, namely the direct integration method and the tanh-method, will be discussed. Both methods seek traveling wave solutions. By working in a traveling frame of reference the PDE is replaced by an ordinary differential equation (ODE) for which one seeks closed-form solutions in terms of special functions.

In the terminology of dynamical systems, the solitary wave solutions correspond to heteroclinic or homoclinic trajectories in the phase plane of a first-order dynamical system corresponding to the underlying ODE (see, e. g., [6]). The periodic solutions are bounded in the phase plane by these special trajectories, which correspond to the limit of infinite period and modulus one.

Other more powerful methods, such as the aforementioned IST and Hirota's method (see, e. g., [40]) deal with the PDE directly. These methods allow one to compute closed-form expressions of soliton solutions (in particular, solitary wave solutions) addressed elsewhere in the encyclopedia.

Direct Integration Method

Exact expressions for solitary wave solutions can be obtained by direct integration. The steps are illustrated for the KdV equation given in (2). Assuming that the wave travels to the right at speed $v = \omega/k$, Eq. (2) can be put into a traveling frame of reference with independent variable $\xi = k(x - vt - x_0)$. This reduces (2) to an ODE, $-v\phi' + \alpha\phi\phi' + k^2\phi''' = 0$, for $\phi(\xi) = u(x,t)$. A first integration with respect to ξ yields

$$-v\phi + \frac{\alpha}{2}\phi^2 + k^2\phi'' = A\,, \tag{28}$$

where A is a constant of integration. Multiplication of (28) by ϕ', followed by a second integration with respect to ξ, yields

$$-\frac{v}{2}\phi^2 + \frac{\alpha}{6}\phi^3 + \frac{k^2}{2}\phi'^2 = A\phi + B\,, \tag{29}$$

where B is an integration constant. Separation of variables and integration then leads to

$$\int_{\phi_0}^{\phi} \frac{d\phi}{\sqrt{a\phi^2 - b\phi^3 + \tilde{A}\phi + \tilde{B}}} = \pm\int_{\xi_0}^{\xi} d\xi\,, \tag{30}$$

where $a = v/k^2$, $b = \alpha/3k^2$, $\tilde{A} = 2A/k^2$, and $\tilde{B} = 2B/k^2$.

The evaluation of the elliptic integral in (30) depends on the relationship between the roots of the function $f(\phi) = a\phi^2 - b\phi^3 + \tilde{A}\phi + \tilde{B}$. In turn, the nature of the roots depends on the choice of $\tilde{A}$ and $\tilde{B}$. Two cases lead to physically relevant solutions.

Case 1: If the three roots are real and distinct, then the integral can be expressed in terms of the inverse of the cn function (see, e. g., [25] for details). This leads to the cnoidal wave solution given in (6).

Case 2: If the three roots are real and (only) two of them coincide, then the tanh-squared solution follows. This happens when $\tilde{A} = \tilde{B} = 0$. Integrating both sides of (30) then gives

$$\begin{aligned}\int_{\phi_0}^{\phi} \frac{d\phi}{\phi\sqrt{a - b\phi}} &= -\frac{2}{\sqrt{a}}\,\mathrm{Arctanh}\left[\frac{\sqrt{a - b\phi}}{\sqrt{a}}\right] + C\\ &= \pm(\xi - \xi_0)\,.\end{aligned} \tag{31}$$

where, without loss of generality, C and ξ_0 can be set to zero. Solving (31) for ϕ yields

$$\phi(\xi) = \frac{a}{b}\left(1 - \tanh^2\left(\frac{\sqrt{a}}{2}\xi\right)\right) = \frac{a}{b}\mathrm{sech}^2\left(\frac{\sqrt{a}}{2}\xi\right). \tag{32}$$

Returning to the original variables, one gets

$$\phi(\xi) = \frac{3v}{\alpha}\mathrm{sech}^2\left(\frac{\sqrt{v}}{2k}\xi\right)\,, \tag{33}$$

or

$$u(x,t) = \frac{3v}{\alpha}\mathrm{sech}^2\left(\frac{\sqrt{v}}{2}(x - vt - x_0)\right)\,, \tag{34}$$

where v is arbitrary. Setting $v = \omega/k = 4k^2$, where k is arbitrary, and $\delta = -kx_0$, one can verify that (34) matches (5).

The Tanh Method

If one is only interested in tanh- or sech-type solutions, one can circumvent explicit integration (often involving elliptic integrals) and apply the so-called tanh-method. A detailed description of the method has been given by Baldwin et al. [5]. The method has been fully implemented in *Mathematica*, a popular symbolic manipulation program, and successfully applied to many nonlinear differential equations from soliton theory and beyond.

The tanh-method is based on the following observation: all derivatives of the tanh function can be expressed as polynomials in tanh. Indeed, using the identity $\cosh^2\xi - \sinh^2\xi = 1$ one computes $\tanh'\xi = \mathrm{sech}^2\xi = 1 - \tanh^2\xi$, $\tanh''\xi = -2\tanh\xi + 2\tanh^3\xi$, etc. Therefore, all derivatives of $T(\xi) = \tanh\xi$ are polynomials in T. For example, $T' = 1 - T^2$, $T'' = -2T + 2T^3$, and $T''' = -2T + 8T^2 - 6T^4$.

By applying the chain rule, the PDE in $u(x,t)$ is then transformed into an ODE for $U(T)$ where $T = \tanh\xi = \tanh(kx - \omega t + \delta)$ is the new independent variable. Since all derivatives of T are polynomials of T, the resulting ODE has polynomial coefficients in T. It is therefore natural to seek a polynomial solution of the ODE. The problem thus becomes algebraic. Indeed, after computing the degree of the polynomial solution, one finds its unknown coefficients by solving a nonlinear algebraic system.

The method is illustrated using (2). Applying the chain rule (repeatedly), the terms of (2) become $u_t = -\omega(1 - T^2)U'$, $u_x = k(1 - T^2)U'$, and

$$\begin{aligned}u_{xxx} = k^3(1 - T^2)\big[&-2(1 - 3T^2)U'\\ &- 6T(1 - T^2)U'' + (1 - T^2)^2U'''\big]\,,\end{aligned} \tag{35}$$

where $U(T) = U(\tanh(kx - \omega t + \delta)) = u(x,t)$, $U' = dU/dT$, etc.

Substitution into (2) and cancellation of a common $1 - T^2$ factor yields

$$\begin{aligned}&-\omega U' + \alpha kUU' - 2k^3(1 - 3T^2)U'\\ &\quad - 6k^3T(1 - T^2)U'' + k^3(1 - T^2)^2U''' = 0\,.\end{aligned} \tag{36}$$

This ODE for $U(T)$ has polynomial coefficients in T. One therefore seeks a polynomial solution

$$U(T) = \sum_{n=0}^{N} a_n T^n \,, \tag{37}$$

where the integer exponent N and the coefficients a_n must be computed.

Substituting T^N into (36) and balancing the highest powers in T gives $N = 2$. Then, substituting

$$U(T) = a_0 + a_1 T + a_2 T^2 \tag{38}$$

into (36), and equating to zero the coefficients of the various power terms in T, yields

$$\begin{aligned} a_1(\alpha a_2 + 2k^2) &= 0 \,, \\ \alpha a_2 + 12k^2 &= 0 \,, \\ a_1(\alpha k a_0 - 2k^3 - \omega) &= 0 \,, \\ \alpha k a_1^2 + 2\alpha k a_0 a_2 - 16k^3 a_2 - 2\omega a_2 &= 0 \,. \end{aligned} \tag{39}$$

The unique solution of this nonlinear system for the unknowns a_0, a_1 and a_2 is

$$a_0 = \frac{8k^3 + \omega}{\alpha k}, \;\; a_1 = 0, \;\; a_2 = -\frac{12k^2}{\alpha} \,. \tag{40}$$

Finally, substituting (40) into (38) and using $T = \tanh(kx - \omega t + \delta)$ yields (4).

The solitary wave solutions and cnoidal wave solutions presented in Sect. "Completely Integrable Shallow Water Wave Equations" have been automatically computed with a *Mathematica* package [5] that implements the tanh-method and variants.

A review of numerical methods to compute solitary waves of arbitrary amplitude can be found in Vanden-Broeck [74].

Water Wave Experiments and Observations

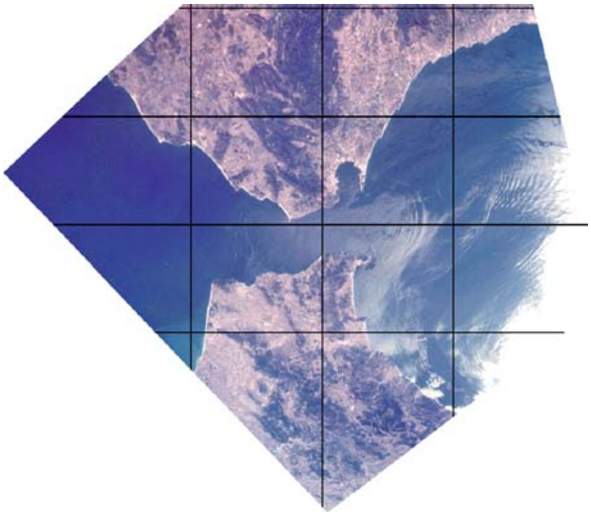
Through a series of experiments in a hydrodynamic tank, Hammack investigated the validity of the BBM equation [35] and KdV equation as models for long waves in shallow water [36,37,38] and long internal waves [69]. Their research addressed the question: Would an initial displacement of water, as it propagates forward, eventually evolve in a train of localized solitary waves (solitons) and an oscillatory tail as predicted by the KdV equation? Based on the experimental data, they concluded that (i) the KdV dynamics only occurs if the waves travel over a long distance, (ii) a substantial amount of water must be initially displaced (by a piston) to produce a soliton train, (iii) the water volume of the initial wave determines the shape of the leading wave in the wave train, and (iv) the initial direction of displacement (upward or downward piston motion) determines what happens later. Quickly raising the piston causes a train of solitons to emerge; quickly lowering it causes all wave energy to distribute into the oscillatory tail, as predicted by the theory.

Several other researchers have tested the validity of the KdV equation and variants in laboratory experiments (see, e. g., [39,60]). Bona et al. [9] give an in-depth evaluation of the BBM Eq. (7) with and without dissipative term u_{xx}. Their study includes (i) a numerical scheme with error estimates, (ii) a convergence test of the computer code, (iii) a comparison between the predictions of the theoretical model and the results of laboratory experiments. The authors note that it is important to include dissipative effects to obtain reasonable agreement between the forecast of the model and the empirical results.

Water tank experiments in conjunction with the analysis of actual data, allows researchers to judge whether or not the KdV equation can be used to model the dynamics of tsunamis (see [67]). Tsunami research intensified after the December 2004 tsunami devastated large coastal regions of India, Indonesia, Sri Lanka, and Thailand, and killed nearly 300,000 people.

Apart from shallow water waves near beaches, the KdV equation and its solitary wave solution also apply to internal waves in the ocean. Internal solitary waves in the open ocean are slow waves of large amplitude that travel at the interface of stratified layers of different density. Stratification based on density differences is primarily due to variations in temperature or concentration (e. g. due to salinity gradients). For example, absorption of solar radiation creates a near surface thin layer of warmer water (of lower density) above a thicker layer or colder, denser water. The smaller the density change, the lower the wave frequency, and the slower the propagation speed. If the upper layer is thinner than the lower one, then the internal wave is a wave of depression causing a downward displacement of the fluid interface.

Internal solitary waves are ubiquitous in stratified waters, in particular, whenever strong tidal currents occur near irregular topography. Such waves have been studied since the 1960s. An early, well-documented case deals with internal waves in the Andaman Sea, where Perry and Schimke [58] found groups of internal waves up to 80 m high and 2000 m long on the main thermocline at 500 m in 1500 m deep water. Their measurements were confirmed by Osborne and Burch [53] who showed that internal waves in the Andaman Sea are generated by tidal flows and can travel over hundreds of kilometers.



Shallow Water Waves and Solitary Waves, Figure 8
Three solitary wave packets generated by internal waves from sills in the Strait of Gibraltar. Original image STS41G-34-81 courtesy of the Earth Sciences and Image Analysis Laboratory, NASA Johnson Space Center (http://eol.jsc.nasa.gov). Ortho-rectified, color adjusted photograph courtesy of Global Ocean Associates

Strong internal waves can affect biological life and interfere with underwater navigation. Understanding the behavior of internal waves can aid in the design of offshore production facilities for oil and natural gas.

The near-surface current associated with the internal wave locally modulates the height of the water surface. Hence, the internal wave leaves a "signature" or "footprint" at the sea surface in the form of a packet of solitary waves (sometimes called current rips or tide rips). These visual manifestations appear as long, quasilinear stripes in satellite imagery or photographs taken during space flights. Over 50 case studies and hundreds of images of oceanic internal waves can be found in "An Atlas of Internal Solitary-like Waves and Their Properties" [42].

Figure 8 shows a photograph of three solitary waves packets which are the surface signature of internal waves in the Strait of Gibraltar. The photograph was taken from the Space Shuttle on October 11, 1984. Spain is to the North, Morocco to the South. Alternate solitary wave packets move toward the northeast or the southeast. The amplitude of these waves is of the order of 50 m; their wavelength is in the range of 500–2000 m. The separation between the packets is approximately 30 km. Waves of longer wavelengths and higher amplitudes have traveled the furthest. The number of oscillations within each packet increases as time goes on. Solitary wave packets can reach 200 km into the Western Mediterranean sea and live for more than two days before dissipating. A in-depth study of solitary waves in the Strait of Gibraltar can be found in Farmer and Armi [28].

The KdV model is applicable to stratified fluids with two layers and internal solitary waves if (i) the ratio of the amplitude a to the upper layer depth h is small, and (ii) the wavelength λ is long compared with the upper layer depth. More precisely, $a/h = O(h^2/\lambda^2) \ll 1$. A detailed discussion of internal solitary waves and additional references can be found in Garrett and Munk [31], Grimshaw [32,33,34], Helfrich and Melville [39], Apel et al. [4], and Pelinovsky et al. [56]. The last three papers discuss a variety of other theoretical models including the extended KdV equation (also known as Gardner's equation or combined KdV-modified KdV equation) which contains both quadratic and cubic nonlinearities. Solitary wave solutions of the extended KdV equation can be found in Scott ([65], p. 856), Helfrich and Melville [39], and Apel et al. [4]. A review of laboratory experiments with internal solitary waves was published by Ostrovsky and Stepanyants [54].

As discussed in the review paper by Staquet and Sommeria [70], internal gravity waves also occur in the atmosphere, where they are often caused by wind blowing over topography and cumulus convective clouds. Internal gravity waves reveal themselves as unusual cloud patterns, which are the counterpart of the solitary wave packets on the ocean's surface.

Future Directions

For many shallow water wave applications, the full Euler equations are too complex to work with. Instead, various approximate models have been proposed. Arguably, the most famous shallow water wave equations are the KdV and Boussinesq equations.

The KdV equation was originally derived to describe shallow water waves in a rectangular channel. Surprisingly, the equation also models ion-acoustic waves and magneto-hydrodynamic waves in plasmas, waves in elastic rods, mid-latitude and equatorial planetary waves, acoustic waves on a crystal lattice, and more (see, e. g., [64,65,66]). The KdV equation has played a pivotal role in the development of the Inverse Scattering Transform and soliton theory, both of which had a lasting impact on twentieth-century mathematical physics.

Historically, the classical Boussinesq equation was derived to describe the propagation of shallow water waves in a canal. Boussinesq systems arise when modeling the propagation of long-crested waves on large bodies of water (e. g. large lakes or the ocean). As Bona et al. [10] point out, a plethora of formally-equivalent Boussinesq systems can be derived. Yet, such systems may have vastly different mathematical properties. The study of the well-posedness of the nonlinear models is of paramount importance and is the subject of ongoing research.

Shallow water wave theory allows one to adequately model waves in canals, surface waves near beaches, and internal waves in the ocean (see [4]). Due to their widespread occurrence in the ocean (see [42]), solitary waves and "solitary wave packets" (solitons) are of interest to oceanographers and geophysicists. The (periodic) cnoidal wave solutions are used by coastal engineers in studies of sediment movement, erosion of sandy beaches, interaction of waves with piers and other coastal structures.

Apart from their physical relevance, the knowledge of solitary and cnoidal wave solutions of nonlinear PDEs facilitates the testing of numerical solvers and aids in stability analysis.

Shallow water wave models are widely used in atmospheric science as a paradigm for geophysical fluid motions. They model, for example, inertia-gravity waves with fast time scale dynamics, and wave-vortex interactions and Rossby waves associated with slow advective-timescale dynamics.

This article has reviewed commonly used shallow water wave models, with the hope of bridging two research communities: one that focuses on nonlinear equations with dispersive effects; the other on nonlinear hyperbolic equations without dispersive terms. Of common concern are the testing of the theoretical models on measured data and further validation of the equations with numerical simulations and laboratory experiments. A fusion of the expertise of both communities might advance research on water waves and help to answer open questions about wave breaking, instability, vorticity, and turbulence. Of paramount importance is the prevention of natural disasters, ecological ravage, and damage to man-made structures due to a better understanding of the dynamics of tsunamis, steep waves, strong internal waves, rips, tidal currents, and storm surges.

Acknowledgments

This material is based upon work supported by the National Science Foundation (NSF) of the USA under Award No. CCF-0830783. The author is grateful to Harvey Segur, Jerry Bona, Michel Grundland, and Douglas Poole for advise and comments. William Navidi, Benoit Huard, Anna Segur, Katherine Socha, Christopher Jackson, Michel Peyrard, and Christopher Eilbeck are thanked for help with figures and photographs.

Bibliography

Primary Literature

1. Ablowitz MJ, Clarkson PA (1991) Solitons, nonlinear evolution equations and inverse scattering. Cambridge University Press, Cambridge
2. Ablowitz MJ, Segur H (1981) Solitons and the inverse scattering transform. SIAM, Philadelphia, Pennsylvania
3. Ablowitz MJ et al (1974) The inverse scattering transform: Fourier analysis for nonlinear problems. Stud Appl Math 53:249–315
4. Apel JR et al (2007) Internal solitons in the ocean and their effect on underwater sound. J Acoust Soc Am 121:695–722
5. Baldwin D et al (2004) Symbolic computation of exact solutions expressible in hyperbolic and elliptic functions for nonlinear PDEs. J Symb Comp 37:669–705
6. Balmforth NJ (1995) Solitary waves and homoclinic orbits. Annu Rev Fluid Mech 27:335–373
7. Benjamin TB et al (1972) Model equations for long waves in nonlinear dispersive systems. Phil Trans Roy Soc London Ser A 272:47–78
8. Bona JL, Chen H (1999) Comparison of model equations for small-amplitude long waves. Nonl Anal 38:625–647
9. Bona JL et al (1981) An evaluation of a model equation for water waves. Phil Trans Roy Soc London Ser A 302:457–510
10. Bona JL et al (2002) Boussinesq equations and other systems for small-amplitude long waves in nonlinear dispersive media. I: Derivation and linear theory. J Nonl Sci 12:283–318
11. Bona JL et al (2004) Boussinesq equations and other systems for small-amplitude long waves in nonlinear dispersive media. II: The nonlinear theory. Nonlinearity 17:925–952
12. Boussinesq J (1871) Théorie de l'intumescence liquide appelée onde solitaire ou de translation, se propageant dans un canal rectangulaire. C R Acad Sci Paris 72:755–759
13. Boussinesq J (1872) Théorie des ondes et des remous qui se propagent le long d'un canal rectangulaire horizontal, en communiquant au liquide contenu dans ce canal des vitesses sensiblement pareilles de la surface au fond. J Math Pures Appl 17:55–108
14. Boussinesq J (1877) Essai sur la théorie des eaux courantes. Académie des Sciences de l'Institut de France, Mémoires présentés par divers savants (ser 2) 23:1–680
15. Bullough RK (1988) "The wave" "par excellence", the solitary, progressive great wave of equilibrium of the fluid–an early history of the solitary wave. In: Lakshmanan M (ed) Solitons: introduction and applications. Springer, Berlin, pp 7–42
16. Camassa R, Holm D (1993) An integrable shallow water equation with peakon solitons. Phys Rev Lett 71:1661–1664
17. Camassa R, Holm D (1994) An new integrable shallow water equation. Adv Appl Mech 31:1–33
18. Chen M (1998) Exact solutions of various Boussinesq systems. Appl Math Lett 11:45–49
19. Clarkson PA, Mansfield EL (1994) On a shallow water wave equation. Nonlinearity 7:975–1000
20. Craik ADD (2004) The origins of water wave theory. Annu Rev Fluid Mech 36:1–28

21. Craik ADD (2005) George Gabriel Stokes and water wave theory. Annu Rev Fluid Mech 37:23–42
22. Darrigol O (2003) The spirited horse, the engineer, and the mathematician: Water waves in nineteenth-century hydrodynamics. Arch Hist Exact Sci 58:21–95
23. Dauxois T, Peyrard M (2004) Physics of solitons. Cambridge University Press, Cambridge, UK. Transl of Peyrard M, Dauxois T (2004) Physique des solitons. Savoirs Actuels, EDP Sciences, CNRS Editions
24. Dellar P (2003) Common Hamiltonian structure of the shallow water equations with horizontal temperature gradients and magnetic fields. Phys Fluids 15:292–297
25. Drazin PG, Johnson RS (1989) Solitons: an introduction. Cambridge University Press, Cambridge, UK
26. El GA (2007) Korteweg–de Vries equation: solitons and undular bores. In: Grimshaw RHJ (ed) Solitary waves in fluids. WIT Press, Boston, pp 19–53
27. Emmerson GS (1977) John Scott Russell: A great Victorian engineer and naval architect. John Murray Publishers, London
28. Farmer DM, Armi L (1988) The flow of Mediterranean water through the Strait of Gibraltar. Prog Oceanogr 21:1–105
29. Fermi F et al (1955) Studies of nonlinear problems I. Los Alamos Sci Lab Rep LA-1940. Reproduced in: AC Newell (ed) (1974) Nonlinear wave motion. AMS, Providence, RI
30. Filippov AT (2000) The versatile soliton. Birkhäuser-Verlag, Basel
31. Garrett C, Munk W (1979) Internal waves in the ocean. Annu Rev Fluid Mech 11:339–369
32. Grimshaw R (1997) Internal solitary waves. In: Liu PLF (ed) Advances in coastal and ocean engineering, vol III. World Scientific, Singapore, pp 1–30
33. Grimshaw R (2001) Internal solitary waves. In: Grimshaw R (ed) Environmental stratified flows. Kluwer, Boston, pp 1–28
34. Grimshaw R (2005) Korteweg–de Vries equation. In: Grimshaw R (ed) Nonlinear waves in fluids: Recent advances and modern applications. Springer, Berlin, pp 1–28
35. Hammack JL (1973) A note on tsunamis: their generation and propagation in an ocean of uniform depth. J Fluid Mech 60:769–800
36. Hammack JL, Segur H (1974) The Korteweg–de Vries equation and water waves, part 2. Comparison with experiments. J Fluid Mech 65:289–314
37. Hammack JL, Segur H (1978a) The Korteweg–de Vries equation and water waves, part 3. Oscillatory waves. J Fluid Mech 84:337–358
38. Hammack JL, Segur H (1978b) Modelling criteria for long water waves. J Fluid Mech 84:359–373
39. Helfrich KR, Melville WK (2006) Long nonlinear internal waves. Annu Rev Fluid Mech 38:395–425
40. Hirota R (2004) The direct method in soliton theory. Cambridge University Press, Cambridge, UK
41. Hirota R, Satsuma J (1976) N-soliton solutions of model equations for shallow water waves. J Phys Soc Jpn 40:611–612
42. Jackson CR (2004) An atlas of internal solitary-like waves and their properties, 2nd edn. Global Ocean Associates, Alexandria, VA
43. Jeffrey A (1978) Nonlinear wave propagation. Z Ang Math Mech (ZAMM) 58:T38–T56
44. Johnson RS (2003) On solutions of the Camassa–Holm equation. Proc Roy Soc London Ser A 459:1687–1708
45. Joseph RI, Egri R (1977) Another possible model equation for long waves in nonlinear dispersive systems. Phys Lett A 61:429–432
46. Kadomtsev BB, Petviashvili VI (1970) On the stability of solitary waves in weakly dispersive media. Sov Phys Doklady 15:539–541
47. Korteweg DJ, de Vries G (1895) On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves. Philos Mag (Ser 5) 39:422–443
48. LeVeque RJ (2002) Finite volume methods for hyperbolic problems. Cambridge University Press, Cambridge, UK
49. Madsen PA, Schäffer HA (1999) A review of Boussinesq-type equations for surface gravity waves. In: Liu PLF (ed) Advances in coastal and ocean engineering, vol V. World Scientific, Singapore, pp 1–95
50. McKean HP (1981) Boussinesq's equation on the circle. Comm Pure Appl Math 34:599–691
51. Miles JW (1981) The Korteweg–de Vries equation: a historical essay. J Fluid Mech 106:131–147
52. Miura RM (1976) The Korteweg–de Vries equation: a survey of results. SIAM Rev 18:412–559
53. Osborne AR, Burch TL (1980) Internal solitons in the Andaman sea. Science 208:451–460
54. Ostrovsky LA, Stepanyants YA (2005) Internal solitons in laboratory experiments: Comparison with theoretical models. Chaos 15:037111
55. Pedlosky J (1987) Geophysical fluid dynamics, 2nd edn. Springer, Berlin
56. Pelinovsky E et al (2007) In: Grimshaw RHJ (ed) Solitary waves in fluids. WIT Press, Boston, pp 85–110
57. Peregrine DH (1966) Calculations of the development of an undular bore. J Fluid Mech 25:321–330
58. Perry RB, Schimke GR (1965) Large amplitude internal waves observed off the north-west coast of Sumatra. J Geophys Res 70:2319–2324
59. Rayleigh L (1876) On waves. Philos Mag 1:257–279
60. Remoissenet M (1999) Waves called solitons: Concepts and experiments, 3rd edn. Springer, Berlin
61. Ripa P (1993) Conservation laws for primitive equations models with inhomogeneous layers. Geophys Astrophys Fluid Dyn 70:85–111
62. Ripa P (1999) On the validity of layered models of ocean dynamics and thermodynamics with reduced vertical resolution. Dyn Atmos Oceans Fluid Dyn 29:1–40
63. Russell JS (1844) Report on waves, 14th meeting of the British Association for the Advancement of Science. John Murray, London, pp 311–390
64. Scott AC (2003) Nonlinear science: Emergence and dynamics of coherent structures. Oxford University Press, Oxford, UK
65. Scott AC (ed) (2005) Encyclopedia of Nonlinear science. Routledge, New York
66. Scott AC et al (1973) The soliton: a new concept in applied science. Proc IEEE 61:1443–1483
67. Segur H (2007a) Waves in shallow water, with emphasis on the tsunami of 2004. In: Kundu A (ed) Tsunami and nonlinear waves. Springer, Berlin
68. Segur H (2007b) Integrable models of waves in shallow water. In: Pinski M, Birnir B (eds) Probability, geometry and integrable systems. Cambridge University Press, Cambridge, UK
69. Segur H, Hammack JL (1982) Soliton models of long internal waves. J Fluid Mech 118:285–304

70. Staquet C, Sommeria J (2002) Internal gravity waves: From instabilities to turbulence. Annu Rev Fluid Mech 34:559–593
71. Stokes GG (1847) On the theory of oscillatory waves. Trans Camb Phil Soc 8:441–455
72. Toro EF (2001) Shock-capturing methods for free-surface shallow flows. Wiley, New York
73. Vallis GK (2006) Atmospheric and oceanic fluid dynamics: fundamentals and large scale circulation. Cambridge University Press, Cambridge, UK
74. Vanden-Broeck J-M (2007) Solitary waves in water: numerical methods and results. In: Grimshaw RHJ (ed) Solitary waves in fluids. WIT Press, Boston, pp 55–84
75. Vreugdenhil CB (1994) Numerical methods for shallow-Water flow. Springer, Berlin
76. Weiyan T (1992) Shallow water hydrodynamics: mathematical theory and numerical solution for two-dimensional systems of shallow water equations. Elsevier, Amsterdam
77. Zabusky NJ (2005) Fermi–Pasta–Ulam, solitons and the fabric of nonlinear and computational science: History, synergetics, and visiometrics. Chaos 15:015102
78. Zabusky NJ, Kruskal MD (1965) Interaction of 'solitons' in a collisionless plasma and the recurrence of initial states. Phys Rev Lett 15:240–243

Books and Reviews

Boyd JP (1998) Weakly nonlinear solitary waves and beyond-all-order asymtotics. Kluwer, Dordrecht

Calogero F, Degasperis A (1982) Spectral transform and solitons I. North Holland, Amsterdam

Dickey LA (2003) Soliton equations and Hamiltonian systems, 2nd edn. World Scientific, Singapore

Dodd RK et al (1982) Solitons and nonlinear wave equations. Academic Press, London

Eilenberger G (1981) Solitons: Mathematical methods for physicists. Springer, Berlin

Faddeev LD, Takhtajan LA (1987) Hamiltonian methods in the theory of solitons. Springer, Berlin

Fordy A (ed) (1990) Soliton theory: A survey of results. Manchester University Press, Manchester

Grimshaw RHJ (ed) (2007) Solitary waves in fluids. WIT Press, Boston

Infeld E, Rowlands G (2000) Nonlinear waves, solitons, and chaos. Cambridge University Press, New York

Johnson RS (1977) A modern introduction to the mathematical theory of water waves. Cambridge University Press, Cambridge, UK

Lamb GL (1980) Elements of soliton theory. Wiley Interscience, New York

Miles JW (1980) Solitary waves. Annu Rev Fluid Mech 12:11–43

Mei CC et al (2005) Theory and applications of ocean surface waves. World Scientific, Singapore

Mitropol'sky YZ (2001) Dynamics of internal gravity waves in the ocean. Kluwer, Dordrecht

Newell AC (1983) The history of the soliton. J Appl Mech 50:1127–1137

Newell AC (1985) Solitons in mathematics and physics. SIAM, Philadelphia, PA

Makhankov VG (1990) Soliton phenomenology. Kluwer, Dordrecht, The Netherlands

Novikov SP et al (1984) Theory of solitons: The inverse scattering method. Consultants Bureau(Plenum Press), New York

Pedlosky J (2003) Waves in the ocean and atmosphere: introduction to wave dynamics. Springer, Berlin

Russell JS (1885) The wave of translation in the oceans of water, air and ether. Trübner, London

Stoker JJ (1957) Water waves. Wiley Interscience, New York

Whitham GB (1974) Linear and nonlinear waves. Wiley Interscience, New York

Signaling Games

Joel Sobel
Department of Economics, University of California,
San Diego, USA

Article Outline

Glossary

Babbling equilibrium An equilibrium in which the sender's strategy is independent of type and the receiver's strategy is independent of signal.

Behavior strategy A strategy for an extensive-form game that specifies the probability of taking each action at each information set.

Behavioral type A player in a game who is constrained to follow a given strategy.

Cheap-talk game A signaling game in which players' preferences do not depend directly on signals.

Condition D1 An equilibrium refinement that requires out-of-equilibrium beliefs to be supported on types that have the most to gain from deviating from a fixed equilibrium.

Divinity An equilibrium refinement that requires out-of-equilibrium beliefs to place relatively more weight on types that gain more from deviating from a fixed equilibrium.

Equilibrium outcome The probability distribution over terminal nodes in a game determined by equilibrium strategy.

Handicap principle The idea that animals communicate fitness through observable characteristics that reduce fitness.

Incomplete information game A game in which players lack information about the strategy sets or payoff functions of their opponents.

Intuitive criterion An equilibrium refinement that requires out-of-equilibrium beliefs to place zero weight on types that can never gain from deviating from a fixed equilibrium outcome.

Nash equilibrium A strategy profile in a game in which each player's strategy is a best response to the equilibrium strategies of the other players.

Neologism-proof equilibrium An equilibrium that admits no self-signaling set.

Pooling equilibrium A signaling-game equilibrium in which each all sender types send the same signal with probability one.

Receiver In a signaling game, the uninformed player.

Self-signaling set A set of types C with the property that precisely types in the set C gain from inducing the best response to C relative to a fixed equilibrium.

Sender In a signaling game, the informed agent.

Separating equilibrium A signaling-game equilibrium in which sender types sent signals from disjoint subsets of the set of available signals.

Signaling game A two-player game of incomplete information in which one player is informed and the other in not. The informed player's strategy is a type-contingent message and the uninformed player's strategy is a message-continent action.

Single-crossing condition A condition that guarantees that indifferent curves from a given family of preferences cross at most one.

Spence-Mirrlees condition A differential condition that orders the slopes of level sets of a function.

Standard signaling game A signaling game in which strategy sets and payoff functions satisfy monotonicity properties.

Type In an incomplete information game, a variable that summarizes private information.

Verifiable information game A signaling game with the property that each type has a signal that can only be sent by that type.

Definition of the Subject

Signaling games refer narrowly to a class of two-player games of incomplete information in which one player is informed and the other is not. The informed player's strategy set consists of signals contingent on information and the uninformed player's strategy set consists of actions contingent on signals. More generally, a signaling game includes any strategic setting in which players can use the actions of their opponents to make inferences about hidden information. The earliest work on signaling games was Spence [72]'s model of educational signaling and Zahari [76]'s model of signaling by animals. During the 1980s researchers developed the formal model and identified conditions that permitted the selection of unique equilibrium outcomes in leading models.

Introduction

The framed degree in your doctor's office, the celebrity endorsement of a popular cosmetic, and the telephone message from an old friend are all signals. The signals are potentially valuable because they allow you to infer useful information. These signals are indirect and require interpretation. They may be subject to manipulation. The doctor's diploma tells you something about the doctor's qualifications, but knowing where and when the doctor studied does not prove that she is a good doctor. The endorsement identifies the product with a particular lifestyle, but what works for the celebrity may not work for you. Besides, the celebrity was probably paid to endorse the product and may not even use it. The phone message may tell you how to get in touch with your friend, but is unlikely to contain all of the information you need to find him – or to evaluate whether you'll meet to discuss old times or to be asked a favor. While the examples all involve signaling, the nature of the signaling is different. The doctor faces large penalties for misrepresenting her credentials. She is not required to display all of her diplomas, but it is reasonable to assume that degrees are not forged. The celebrity endorsement is costly – certainly to the manufacturer who pays for the celebrity's services and possibly to the celebrity himself, whose reputation may suffer if the product works badly. It is reasonable to assume that it is easier to obtain an endorsement of a good product, but there are also good reasons to be skeptical about the claims. In contrast, although a dishonest or misleading message may lead to a bad outcome, leaving a message is not expensive and the content of the message is not constrained by your friend's information. The theory of signaling games is a useful way to describe the essential features of all three examples.

Opportunities to send and evaluate signals arise in many common natural and economic settings. In the canonical example (due to Spence [72]), a high-ability worker invests in education to distinguish herself from less skilled workers. The potential employer observers educational attainment, but not innate skill, and infers that a better educated worker is more highly skilled and pays a higher wage. To make this story work, there must be a reason that low-ability workers do not get the education expected of a more highly skilled worker and hence obtain a higher wage. This property follows from an assumption that the higher the ability the worker, the easier it is for her to produce a higher signal.

The same argument appears in many applications. For example, a risk-averse driver will purchase a lower cost, partial insurance contract, leaving the riskier driver to pay a higher rate for full insurance (Rothschild and Stiglitz [65] or Wilson [75]). A firm that is able to produce high-quality goods signals this by offering a warranty for the goods sold (Grossman [37]) or advertising extensively. A strong deer grows extra large antlers to show that it can survive with this handicap and to signal its fitness to potential mates [76].

Game theory provides a formal language to study how one should send and interpret signals in strategic environments. This article reviews the basic theory of signaling and discusses some applications. It does not discuss related models of screening. Kreps and Sobel [43] and Riley [64] review both signaling and screening.

Section "The Model" describes the basic model. Section "Equilibrium" defines equilibrium for the basic model. Section "The Basic Model" limits attention to a special class of signaling game. I give conditions sufficient for the existence of equilibria in which the informed agent's signal fully reveals her private information and argue that one equilibrium of this kind is prominent. The next three sections study different signaling games. Section "Cheap Talk" discusses models of costless communication. Section "Verifiable Information" discusses the implications of the assumptions that some information is verifiable. Section "Communication about Intentions" briefly discusses the possibility of signaling intentions rather than private information. Section "Applications" describes some applications and extensions of the basic model. Section "Future Directions" speculates on directions for future research.

The Model

This section describes the basic signaling game. There are two players, called S (for sender) and R (for receiver). S knows the value of some random variable t whose support is a given set T. t is called the **type** of S. The prior beliefs of R are given by a probability distribution $\pi(\cdot)$ over T; these beliefs are common knowledge. When T is finite, $\pi(t)$ is the prior probability that the sender's type is t. When T is uncountably infinite, $\pi(\cdot)$ is a density function. Player S learns t and sends to R a signal s drawn from some set M. Player R receives this signal, and then takes an action a drawn from a set A. (It is possible to allow A to depend on s and S to depend on t.) This ends the game: The payoff to i is given by a function $u^i : T \times M \times A \to \mathbb{R}$.

This canonical game captures the essential features of the classic applications of market signaling. In the labor-market signaling story due to Spence [72] a worker wishes to signal his ability to a potential employer. The worker has information about ability that the employer lacks. Direct communication about ability is not possible, but the worker can acquire education. The employer can observe the worker's level of education and use this to form a judgment about the worker's true level of ability. In this application, S is a worker; R represents a potential employer (or a competitive labor market); t is the student's productivity; s is her level of education; and a is her wage.

Equilibrium

Defining Nash equilibrium for the basic signaling game is completely straightforward when T, S, and A are finite sets. In this case a behavior strategy for S is a function $\mu : T \times M \to [0, 1]$ such that $\sum_{s \in M} \mu(t, s) = 1$ for all t. $\mu(t, s)$ is the probability that sender-type t sends the signal s. A behavior strategy for R is a function $\alpha : M \times A \to [0, 1]$ where $\sum_{a \in A} \alpha(s, a) = 1$ for all s. $\alpha(s, a)$ is the probability that R takes action a following the signal s.

Proposition 1 *Behavior strategies (α^*, μ^*) form a Nash Equilibrium if and only if for all $t \in T$*

$$\mu(t,s) > 0 \ \textit{implies} \ \sum_{a \in A} U^S(t,s,a)\alpha(s,a) = \max_{s' \in S} \sum_{a \in A} U^S(t,s',a)\alpha(s',a) \quad (1)$$

and, for each $s \in S$ such that $\sum_{t \in T} \mu(t,s)\pi(t) > 0$, if $\sum_{t \in T} \mu(t,s)\pi(t) > 0$, then

$$\alpha(s,a) > 0 \ \textit{implies} \ \sum_{t \in T} U^R(t,s,a)\beta(t,a) = \max_{a' \in A} \sum_{t \in T} U^R(t,s,a')\beta(t,a')\,, \quad (2)$$

where

$$\beta(t,s) = \frac{\mu(t,s)\pi(t)}{\sum_{t' \in T} \mu(t',s)\pi(t')}\,. \quad (3)$$

Condition (1) states that the S places positive probability only on signals that maximize expected utility. This condition guarantees that S responds optimally to R's strategy.

Condition (2) states that R places positive probability only on actions that maximize expected utility, where is taken with respect to the distribution $\beta(\cdot, s)$ following the signal s. Condition (3) states that $\beta(\cdot, s)$ accurately reflects the pattern of play. It requires that R's beliefs be determined using S's strategy and the prior distribution whenever possible. Equilibrium refinements also require that R has beliefs following signals s that satisfy

$$\sum_{t \in T} \mu(t, s)\pi(t) = 0\,, \tag{4}$$

that is are sent with probability zero in equilibrium. Specifically, sequential equilibrium permits $\beta(\cdot, m)$ to be an arbitrary distribution when Eq. (4) holds, but requires that Eq. (2) holds even for these values of s. This restriction rules out equilibria in which certain signals are not sent because the receiver responds to the signal with an action that is dominated.

The ability to signal creates the possibility that R will be able to draw inferences about S's type from the signal. Whether he is able to do so is a property of the equilibrium. It is useful to define two extreme cases.

Definition 1 An equilibrium (α^*, μ^*) is called a separating equilibrium if each type t sends different signals. That is, M can be partitioned into sets M_t such that for each t, $\sum_{s \in M_t} \mu(t, s) = 1$. An equilibrium (α^*, μ^*) is called a pooling equilibrium if there is a single signal s^* that is sent by all types with probability one.

In a separating equilibrium, R can infer S's private information completely. In a pooling equilibrium, R learns nothing from the sender's signal. This definition excludes other possible situations. For example, all sender types can randomize uniformly over a set of two or more signals. In this case, the receiver will be able to draw no inference beyond the prior from a signal received in equilibrium. More interesting is the possibility that the equilibrium will be partially revealing, with some, but not all of the sender types sending common signals.

It is not difficult to construct pooling equilibria for the basic signaling game. Take the labor market model and assume S sends the message s^* with probability one and that the receiver responds to s^* with his best response to the prior distribution and to all other messages with the best response to the belief that t is the least skilled agent. Provided that the least skilled agent prefers to send s^* to sending the cheapest alternative signal, this is a Nash Equilibrium outcome.

The Basic Model

The separating equilibrium is a benchmark outcome for signaling games. When a separating equilibrium exists, then it is possible for the sender to share her information fully with the receiver in spite of having a potential conflict of interest.

Existence of separating equilibria typically requires a systematic relationship between types and signals. An appropriate condition, commonly referred to as the single-crossing condition, plays a prominent role in signaling games and in models of asymmetric information more generally.

In this section I limit attention to a special class of signaling game in which there is a monotonic relationship between types and signals. In these models, separating equilibria typically exist.

I begin by stating the assumption in the environment most commonly seen in applications. Assume that the sets T, S, and A are all real intervals.

Definition 2 $U^S(\cdot)$ satisfies the single-crossing condition if $U^S(t, s, a) \le U^S(t, s', a')$ for $s' > s$ implies that $U^S(t', s, a) < U^S(t', s', a')$ for all $t' > t$.

In a typical application, $U^S(\cdot)$ is strictly decreasing in its second argument (the signal) and increasing in its third argument (R's response) for all types. Consequently indifference curves are well defined in $M \times A$ for all t. The single-crossing condition states that indifference curves of different sender types cross once. If a lower type is indifferent between two signal-action pairs, then a higher type strictly prefers to send the higher signal. In this way, the single-crossing condition links signals to types in such a way as to guarantee that higher types send weakly higher signals in equilibrium.

Note two generalizations of Definition 2. First, the assumption that the domain of $U^S(\cdot)$ is the product of intervals can be replaced by the assumption that these sets are partially ordered. In this case, weak and strict order replace the weak and strict inequalities comparing types and actions in the statement of the definition. Second, it is sometimes necessary to extend the definition to mixed strategies. In this case, the ordering of A induces a partial ordering of distributions of A through first-order stochastic dominance.

When one thinks of the single-crossing condition geometrically, it is apparent that it implies a ranking of the slopes of the indifference curves of the sender. Suppose that $U^S(\cdot)$ is smooth, strictly increasing in actions and strictly decreasing in signals so that indifference curves are well defined for each t. Writing the indifference curve as $\{(s, \bar{a}(s; t))\}$, it must be that $U^S(t, s, \bar{a}(s; t)) \equiv 0$, so that the slope of the indifference curve of a type t sender is

$$\bar{a}_1(s;t) = -\frac{U_2^S(t,s,a)}{U_3^S(t,s,a)}\,, \tag{5}$$

where $\bar{a}_1(s;t)$ is the partial derivative of $\bar{a}(s;t)$ with respect to the first argument, and $U_k^S(\cdot)$ denotes the partial derivative of $U^S(\cdot)$ with respect to its kth argument. Under these conditions, the single-crossing condition is implied by the requirement that the right-hand side of Eq. (5) is decreasing in t. The differentiable version of the single-crossing condition is often referred to as the Spence–Mirrlees condition. Milgrom and Shannon [57] contains general definitions of the single-crossing and Spence–Mirrlees conditions and Edlin and Shannon [26] provides a precise statement of when the conditions are equivalent.

To provide a simple construction of a separating equilibrium, limit attention to a standard signaling game in which the following conditions hold.

1. $T = \{0, \ldots, K\}$ is finite.
2. A and M are real intervals.
3. Utility functions are continuous in action and signal.
4. $U^S(\cdot)$ is strictly increasing in action and strictly decreasing in signal.
5. The single-crossing property holds.
6. The receiver's best-response function is uniquely defined, independent of the signal, and strictly increasing in t so that it can be written $BR(t)$.
7. There exists $\bar{s} \in S$ such that $U^S(K, \bar{s}, BR(K)) < U^S(K, s_0^*, BR(0))$.

Conditions 1 and 2 simplify exposition, but otherwise are not necessary. It is important that T, M, and A be partially ordered so that some kind of single-crossing condition applies. Conditions 4–6 impose a monotone structure on the problem so that higher types are more able to send high signals, and that higher types induce higher (and uniformly more attractive) actions. These conditions imply that in equilibrium higher types will necessarily send weakly higher signals. Condition 7 is a boundary condition that makes sending high signals unattractive. It states that the highest type of sender would prefer to be treated like the lowest type rather than use the signal $\bar{s}$. These properties hold in many standard applications. Condition 6 would be satisfied if $U^R(t,s,a) = -(a-t)^2$.

Separating Equilibrium

To illustrate these ideas, consider a construction of a separating equilibrium.

Proposition 2 *The standard signaling game has a separating equilibrium.*

One can prove the proposition by constructing a possible equilibrium path and confirming that the path can be part of a separating equilibrium.

Step 1. t_0 selects the signal s_0^* that maximizes $U^S(t_0, s, BR(t_0))$.

Step 2. Suppose that s_i^* have been specified for $i = 0, \ldots, k-1$ and let $U^*(t_i) = U^S(t_i, s_i^*, BR(t_i))$. Define s_k^* to solve:

$$\max U^S(t_k, s, BR(t_k)) \text{ subject to } U^S(t_{k-1}, s, BR(t_k)) \leq U^*(t_{k-1})\,.$$

Provided that the optimization problems in Steps 1 and 2 have solutions, the process inductively produces a signaling strategy for the sender and a response rule for the receiver defined on $\{s_0^*, \ldots, s_K^*\}$. When $BR(\cdot)$ is strictly increasing, the single-crossing condition implies that the signaling strategy is strictly increasing. To complete the description of strategies, assume that the receiver takes the action $BR(t_k)$ in response to signals in the interval $[s_k, s_{k+1})$, $BR(t_0)$ for $s < s_0^*$, and $BR(t_K)$ for $s > s_K^*$. By the definition of the best-response function, the receiver is best responding to the sender's strategy. When the boundary condition fails, a fully separating equilibrium need not exist, but when M is compact, one can follow the construction above to obtain an equilibrium in which the lowest types separate and higher types pool at the maximum signal in M (see Cho and Sobel [22] for details).

In the construction, the equilibrium involves inefficient levels of signaling. When $U^S(\cdot)$ is decreasing in the signal, all but the lowest type of sender must make a wasteful expenditure in the signal in order to avoid being treating as having a lower quality. The result that expenditures on signals are greater than the levels optimal in a full-information model continue to hold when $U^S(\cdot)$ is not monotonic in the signal. The sender inevitably does no better in a separating equilibrium than she would do if R had full information about t. Indeed, all but the lowest type will do strictly worse in standard signaling games. On the other hand, the equilibrium constructed above has a constrained efficiency property: Of all separating equilibria, it is Pareto dominant from the standpoint of S. To confirm this claim argue inductively that in any separating equilibrium if t_j sends the signal s_j, then $s_j \geq s_j^*$, with equality only if all types $i < j$ send s_i^* with probability one.

Mailath [49] provides a similar construction when T is a real interval. In this case, the Spence–Mirrlees formulation of the single-crossing condition plays an important role and the equilibrium is a solution to a differential equation.

Multiple Equilibria and Selection

Section "Equilibrium" ended with the construction of a pooling equilibrium. A careful reconsideration of the argument reveals that there typically are many pooling equilibrium outcomes. One can construct a potential pooling outcome by assuming that all sender types send the same signal, that the receiver best responds to this common signal, and responds to all other signals with the least attractive action. Under the standard monotonicity assumptions, this strategy profile will be an equilibrium if the lowest sender type prefers pooling to sending the cheapest available out-of-equilibrium message. Section "Separating Equilibrium" ended with the construction of a separating equilibrium. There are also typically many separating equilibrium outcomes. Assume that types $t = 0, \dots, r-1$ send signals $s^*(t)$, type r sends $\tilde{s}(k) > s^*(k)$, and subsequent signals $\tilde{s}^*(t)$ for $t > r$ solve:

$$\max U^S(t_k, s, BR(t_k)) \text{ subject to}$$
$$U^S(t_{k-1}, s, BR(t_k)) \le U(t_{k-1}, \tilde{s}, BR(t_{k-1})) .$$

In both of these cases, the multiplicity is typically profound, with a continuum of distinct equilibrium outcomes (when M is an interval). The multiplicity of equilibria means that, without refinement, equilibrium theory provides few clear predictions beyond the observation that the lowest type of sender receives at least $U^*(t_0)$, the payoff it would receive under complete information, and the fact that the equilibrium signaling function is weakly increasing in the sender's type. The first property is a consequence of the monotonicity of S's payoff in a and of R's best response function. The second property is a consequence of the single-crossing condition.

This section describes techniques that refine the set of equilibria. Refinement arguments that guarantee existence and select unique outcomes for standard signaling games rely on the Kohlberg–Mertens [42] notion of strategic stability. The complete theory of strategic stability is only available for finite games. Consequently the literature applies weaker versions of strategic stability that are defined more easily for large games. Banks and Sobel [8], Cho and Kreps [21], and Cho and Sobel [22] introduce these arguments.

Multiple equilibria arise in signaling games because Nash equilibrium does not constrain the receiver's response to signals sent with zero probability in equilibrium. Specifying that R's response to these unsent signals is unattractive leads to the largest set of equilibrium outcomes. (In standard signaling games, S's preferences over actions do not depend on type, so the least attractive action is well defined.) The equilibrium set shrinks if one restricts the meaning of unsent signals. An effective restriction is condition D1, introduced in Cho and Kreps [21]. This condition is less restrictive than the notion of universal divinity introduced by Banks and Sobel [8], which in finite games is less restrictive than Kohlberg and Mertens's notion of strategic stability.

Given an equilibrium (α^*, μ^*), let $U^*(t)$ be the equilibrium expected payoff of a type t sender and let $D(s,t) = \{a \colon u(t,s,a) \ge U^*(t)\}$ be the set of pure-strategy responses to s that lead to payoffs at least as great as the equilibrium payoff for player t. Given a collection of sets, $X(t), t \in T$, $X(t^*)$ is maximal if it not a proper subset of any $X(t)$.

Definition 3 Behavior strategies (α^*, μ^*) together with beliefs β^* satisfy D1 if for any unsent message s, $\beta(\cdot, s)$ is supported on those t for which $D(s,t)$ is maximal.

In standard signaling games, $D(s,t)$ is an interval: all actions greater than or equal to a particular action will be attractive relative to the equilibrium. Hence these sets are nested. If $D(s,t)$ is not maximal, then there is another type t' that is "more likely to deviate" in the sense that there exists out-of-equilibrium responses that are attractive to t' but not t. Condition D1 requires that the receiver place no weight on type t making a deviation in this case. Notice if $D(s,t)$ is empty for all t, then D1 does not restrict beliefs given s (and any choice of action will support the putative equilibrium). Condition D1 is strong. One can imagine weaker restrictions. The intuitive condition (Cho and Kreps [21]) requires that $\beta(t,s) = 0$ when $D(t,s) = \phi$ and at least one other $D(t',s)$ is non empty. Divinity (Banks and Sobel [8]) requires that if $D(t,s)$ is strictly contained in $D(t',s)$, then $\beta(t',s)/\beta(t,s) \ge \pi(t')/\pi(t)$, so that the relative probability of the types more likely to deviate increases.

Proposition 3 *The standard signaling game has a unique separating equilibrium outcome that satisfies Condition D1.*

In standard signaling games, the only equilibrium outcome that satisfies Condition D1 is the separating outcome described in the previous section. Details of the argument appear in Cho and Sobel. The argument relies on two insights. First, types cannot be pooled in equilibrium because slightly higher signals will be interpreted as coming from the highest type in the pool. Second, in any separating equilibrium in which a sender type fails to solve Step 2, deviation to a slightly lower signal will not lower R's beliefs.

The refinement argument is powerful and the separating outcome selected receives prominent attention in the literature. It is worth pointing out that the outcome has one unreasonable property. The separating outcome described above depends only on the support of types, and not on the details of the distribution. Further, all types but the lowest type must make inefficient (compared to the full-information case) investments in signal in order to distinguish themselves from lower types. The efficient separating equilibrium for a sequence of games in which the probability of the lowest type converges to zero does not converge to the separating equilibrium of the game in which the probability of the lowest type is zero. In the special case of only two types, the (efficient) pooling outcome may be a more plausible outcome when the probability of the lower type shrinks to zero. Grossman and Perry [38] and Mailath, Okuno-Fujiwara, and Postlewaite [50] introduce equilibrium refinements that select the pooling equilibrium in this setting. These concepts share many of the same motivations of the refinements introduced by Banks and Sobel and Cho and Kreps. They are qualitatively different from the intuitive criterion, divinity, and Condition D1, because they are not based on dominance arguments and lack general existence properties.

Cheap Talk

Models in which preferences satisfy the single-crossing property are central in the literature, but the assumption is not appropriate in some interesting settings. This section describes an extreme case in which there is no direct cost of signaling.

In general, a cheap-talk model is a signaling model in which $u^i(t, s, a)$ is independent of s for all (t, a). Two facts about this model are immediate. First, if equilibrium exists, then there always exists an equilibrium in which no information is communicated. To construct this "babbling" equilibrium, assume that $\beta(t, s)$ is equal to the prior independent of the signal s. R's best response will be to take an action that is optimal conditional only on his prior information. Hence R's action can be taken to be constant. In this case, it is also a best response for S to send a signal that is independent of type, which makes $\beta(t, s)$ the appropriate beliefs. Hence, even if the interests of S and R are identical, so that it there are strong incentives to communicate, there is a possibility of complete communication break down.

Second, it is clear that non-trivial communication requires that different types of S have different preferences over R's actions. If it is the case that whenever some type t prefers action a to action a' then so do all other types, then (ruling out indifference), it must be the case that in equilibrium the receiver takes only one action with positive probability. To see this, note that otherwise one type of sender is not selecting a best response. The second observation shows that cheap talk is not effective in games, like the standard labor-market story, in which the sender's preferences are monotonic in the action of the receiver. With cheap communication, the potential employee in the labor market will always select a signal that leads to the higher possible wage and consequently, in equilibrium, all types of workers will receive the same wage.

A Simple Cheap-Talk Game

There are natural settings in which cheap talk is meaningful in equilibrium. To describe examples, I follow the development of Crawford and Sobel [24] (Green and Stokey [35] independently introduced a similar game in an article circulated in 1981). In this paper, A and T are the unit interval and M can be taken to be the unit interval without loss of generality. The sender's private information or type, t, is drawn from a differentiable probability distribution function, $F(\cdot)$, with density $f(\cdot)$, supported on $[0, 1]$. S and R have twice continuously differentiable von Neumann–Morgenstern utility functions $U^i(a, t)$ that are strictly concave in a and have a strictly positive mixed partial derivative. Let $i = R, S, a^i(t)$ denotes the unique solution to $\max_a U^i(a, t)$ and further assume that $a^S(t) > a^R(t)$ for all t. (The assumptions on $U^i(\cdot)$ guarantee that $U^i(\cdot)$ is well defined and strictly increasing.)

In this model, the interests of the sender and receiver are partially aligned because both would like to take a higher action with a higher t. The interests are different because S would always like the action to be a bit higher than R's ideal action. In a typical application, t represents the idea action for R, such as the appropriate expenditure on a public project. Both R and S want actual expenditure to be close to the target value, but S has a bias in favor of additional expenditure.

For $0 \le t' < t'' \le 1$, let $\bar{a}(t', t'')$ be the unique solution to $\max_a \int_{t'}^{t''} U^R(a, t)\mathrm{d}F(t)$. By convention, $\bar{a}(t, t) = a^R(t)$.

Without loss of generality, limit attention to pure-strategy equilibria. The concavity assumption guarantees that R's best responses will be unique, so R will not randomize in equilibrium. An equilibrium with strategies (μ^*, α^*) **induces** action a if $\{t\colon \alpha^*(\mu^*(t)) = a\}$ has positive prior probability. Crawford and Sobel [24] characterize equilibrium outcomes.

Proposition 4 *There exists a positive integer N^* such that for every integer N with $1 \le N \le N^*$, there exists at least one equilibrium in which the set of induced actions has cardinality N, and moreover, there is no equilibrium* *which induces more than N^* actions. An equilibrium can be characterized by a partition of the set of types, $t(N) = (t_0(N), \ldots, t_N(N))$ with $0 = t_0(N) < t_1(N) < \ldots < t_N(N) = 1$, and signals m_i, $i = 1, \ldots, N$, such that for all $i = 1, \ldots, N-1$*

$$U^S(\bar{a}(t_i, t_{i+1}), t_i)) - U^S(\bar{a}(t_{i-1}, t_i), t_i)) = 0\,, \tag{6}$$

$$\mu(t) = m_i \ \text{for} \ \ t \in (t_{i-1}, t_i]\,, \tag{7}$$

and

$$\alpha(m_i) = \bar{a}(t_{i-1}, t_i)\,. \tag{8}$$

Furthermore, essentially all equilibrium outcomes can be described in this way.

In an equilibrium, adjacent types pool together and send a common message. Condition (6) states that sender types on the boundary of a partition element are indifferent between pooling with types immediately below or immediately above. Condition (7) states that types in a common element of the partition send the same message. Condition (8) states that R best responds to the information in S's message.

Crawford and Sobel make another monotonicity assumption, which they call condition (M). (M) is satisfied in leading examples and implies that there is a unique equilibrium partition for each $N = 1, \ldots, N^*$, the ex-ante equilibrium expected utility for both S and R is increasing in N, and N^* increases if the preferences of S and R become more aligned. These conclusions provide justification for the view that with fixed preferences "more" communication (in the sense of more actions induced) is better for both players and that the closer are the interests of the players the greater the possibilities for communication.

As in the case of models with costly signaling, there are multiple equilibria in the cheap-talk model. The multiplicity is qualitatively different. Costly signaling models have a continuum of Nash Equilibrium outcomes. Cheap-talk models have only finitely many. Refinements that impose restrictions on off-the-equilibrium path signals work well to identify a single outcome in costly signaling models. These refinements have no cutting power in cheap-talk models because any equilibrium distribution on type-action pairs can arise from signaling strategies in which all messages are sent with positive probability. To prove this claim, observe that if message m' is unused in equilibrium, while message m is used, then one can construct a new equilibrium in which R interprets m' the same way as m and sender types previously sending m randomize equally between m and m'.

In the basic model messages take on meaning only through their use in an equilibrium. Unlike natural language, they have no external meaning. There have been several attempts to formalize the notion that messages have meanings that, if consistent with strategic aspects of the interaction, should be their interpretation inside the game. The first formulation of this idea is due to Farrell [28].

Definition 4 Given an equilibrium (α^*, σ^*) with sender expected payoffs $u^*(\cdot)$, the subset $G \subset T$ is self signaling if $G = \{t \colon U^S(t, BR(G)) > u^*(t)\}$.

That is, G is self signaling if precisely the types in G gain by making a statement that induces the action that is a best response to the information that $t \in G$. (When $BR(t)$ is not single valued it is necessary to refine the definition somewhat and permit the possibility that $U^S(t, BR(G)) = u^*(t)$ for some t. See Matthews, Okuno-Fujiwara, and Postlewaite [51].) Farrell argues that the existence of a self-signaling set would destroy an equilibrium. If a subset G had available a message that meant "my type is in G," then relative to the equilibrium R could infer that if he were to interpret the message literally, then it would be sent only by those types in G (and hence the literal meaning would be accurate). With this motivation, Farrell proposes a refinement.

Definition 5 An equilibrium (α^*, σ^*) is neologism proof if there exist no self-signaling sets relative to the equilibrium.

Rabin [62] argues convincingly that Farrell's definition rules out too many equilibrium outcomes. Indeed, for leading examples of the basic cheap-talk game, there are no neologism-proof equilibria. Specifically, in the Crawford–Sobel model in which S has a bias towards higher actions, there exist self signaling sets of the form $[t, 1]$. On the other hand, Chen, Kartik, and Sobel [20] demonstrate that the N^*-step equilibrium always satisfies the no incentive to separate (NITS) condition:

$$U^S(\alpha^*(\mu^*(0)), 0) \geq U^S(a^R(0), 0)\,, \tag{9}$$

and that under condition (M) this is the only equilibrium that satisfies Condition (9).

NITS states that the lowest type of sender prefers her equilibrium payoff to the payoff she would receive if the receiver knew her type (and responded optimally). [41] introduced and named this condition. The NITS condition can be shown to rule out equilibria that admit if self-signaling sets of the form $[0, t]$. Chen [19] and Kartik [41] show that the condition holds in the limits of perturbed versions of the basic cheap-talk game.

Inequality (9) holds automatically in any perfect Bayesian equilibrium of the standard signaling model. This follows because when R's actions are monotonic in type and S's preferences are monotonic in action, the worst outcome for S is to be viewed as the lowest type. This observation would not be true in Nash Equilibrium, where it is possible for R to respond to an out-of-equilibirum message with an action $a < BR(0)$.

Variations on Cheap Talk

In standard signaling models, there is typically an equilibrium that is fully revealing. This is not the case in the basic cheap-talk model. This leads to the question of whether it is possible to obtain more revelation in different environments.

One possibility is to consider the possibility of signaling over many dimensions. Chakraborty and Harbaugh [18] consider a model in with T and A are multidimensional. A special case of their model is one in which the components of T are independent draws from the same distribution and A involves taking a real-valued action for each component of T. If preferences are additively separable across types and actions, Chakraborty and Harbaugh provide conditions under which categorical information transmission, in which the S transmits the order of the components of T, is credible in equilibrium even when it would not be possible to transmit information if the dimensions were treated in isolation. It may be credible for S to say "$t_1 > t_2$," even if she could not credibly provide information about the absolute value of either component of t.

Communication is non-trivial if some Sender type strictly prefers to induce one equilibrium action over another. Non-trivial communication requires that different types have different preferences over outcomes. In standard signaling models, the heterogeneity arises because different sender types have different costs of sending messages. In cheap-talk models, the heterogeneity arises with one-dimensional actions if different sender types have different ideal actions. With multi-dimensional actions, heterogeneity could come simply from different sender types having different preferences over the relative importance of the different issues. Another simple variation is to assume the existence of more than one sender. In the two-sender game, nature picks t as before, both senders learn t and simultaneously send a message to the receiver, who makes a decision based on the two messages. The second sender has preferences that depend on type and the receiver's action, but not directly on the message sent. In this environment, assume that $M = T$, so that the set of available messages (this is essentially without loss of generality). One can look for equilibria in which the senders report honestly. Denote by $a^*(t, t')$ R's response to the pair of messages (t, t'). If an equilibrium in which both senders report honestly exists, then R's response to identical messages, $a^*(t, t) = a^R(t)$, and it must be the case that there exists a specification of $a(t, t')$ for $t \neq t'$ such that for all $i = 1$ and 2 and $t \neq t'$,

$$U^{S_i}(t, a^*(t, t)) \geq U^{S_i}(t, a^*(t, t')) \,. \tag{10}$$

It is possible to satisfy Condition (10) if the biases of the senders are small relative to the set of possible best responses. Krishna and Morgan [45] studies a one-dimensional model of information transmission with two informed players. Ambrus and Takahashi [1] and Battiglini [9] provide conditions under which full revelation is possible when there are two informed players and possibly multiple dimensions of information.

In many circumstances, enriching the communication structure either by allowing more rounds of communication [2,29], mediation [10], or exogenous uncertainty [16] enlarges the set of equilibrium outcomes.

Verifiable Information

Until now, the focus has been on situations in which the set of signals available does not depend on the true state. There are situations in which this assumption is not appropriate. There may be laws that ban false advertisement. The sender may be able to document details about the value of t. Models of this kind were first studied by Grossman [37] and Milgrom [56]. For example, if t is the sender's skill at playing the piano, then if there is a piano available t could demonstrate that she has skill at least as great as t (by performing at her true ability), but she may not be able to prove that her skill is no more than t (the receiver may think that she deliberately played the piano badly).

To model these possibilities, suppose that the set of possible messages is the set of all subsets of T. In this case, messages have "literal" meanings: When the sender uses the message $s = C \in T$, this can be interpreted as a statement of the form: "my type is in C." If senders cannot lie, then $M(t)$ must be the set of subsets of T that contain t. If type t is verifiable, then $\{t\} \in M(t')$ if and only if $t' = t$. If there are no additional costs of sending signals, this model can be viewed as a variation of cheap talk models in which the message space depends on t. In general, one can treat verifiable information models as a special case of the general signaling game in which the cost of sending certain signals is so large that these signals can be ruled out. Lying is impossible if $M(t) = \{C \subset 2^T : t \in C\}$. In this setting, it is appropriate to require equilibria to be consistent with the signaling structure.

Definition 6 The equilibrium (σ^*, α^*) is rationalizable if

$$\alpha(C, a) > 0 \text{ implies } \sum_{t \in T} U^R(t, s, a)\beta(t, a) = \max_{a' \in A} \sum_{t \in T} U^R(t, s, a')\beta(t, a') , \quad (11)$$

where $\beta(t, a) = 0$ if $t \notin C$.

Compared to (2), (11) requires that beliefs place positive probability only on types capable of sending the message "my type is an element of C."

Proposition 5 *Suppose that A and T are linearly ordered, that the receiver's best response function is increasing in type, and that all sender types prefer higher actions. If lying is not possible, then in any rationalizable equilibrium* (α^*, μ^*) , $\alpha^*(s, BR(t)) = 1$ *whenever* $\mu^*(t, s > 0)$.

Grossman [37] and Milgrom [56] present versions of this proposition. Seidman and Winter [68] generalize the result.

Provided that the receiver responds to the signal $\{t\}$ with $BR(t)$, each type can guarantee a payoff of $BR(t)$. On the other hand, if any type receives a payoff greater than $BR(t)$, then some higher type must be doing worse. Another way to make the same point is to notice that the highest type $\{\bar{t}\}$ has a weakly dominant strategy to reveal her type by announcing $\{\bar{t}\}$. Once this type is revealed, the next highest type will want to reveal herself and so on. Hence verifiable information will be revealed voluntarily in an environment where cheap talk leads to no revealing and costly signaling will be compatible with full revelation, but at the cost of dissipative signaling.

The full-revelation result depends on the assumption that the sender and receiver share a linear ranking over the quality of information. Giovannini and Seidmann [31] discuss more general settings in which the ability to provide verifiable information need not lead to full revelation.

Communication About Intentions

In a simple signaling game, signals potentially provide information about private information. Another possibility is to add a round of pre-play communication to a given game. Even if the game has complete information, there is the possibility that communication would serve to select equilibria or permit correlation that would otherwise be infeasible. Farrell and Rabin [63]'s review article discusses this literature in more detail.

Aumann [3] argues that one cannot rely on pre-play communication to select a Pareto-efficient equilibrium. He considers a simple two-player game with Pareto-ranked equilibria and argues that no "cheap" pre-play signal would be credible.

Ben-Porath and Dekel [11] show that adding a stage of "money burning" (a signal that reduces all future payoffs by the same amount) when combined with an equilibrium refinement can select equilibria in a complete information game. Although no money is burned in the selected equilibrium outcome, the potential to send costly signals creates dominance relationships that lead to a selection.

Vida [74] synthesizes a literature that compares the set of equilibrium outcomes available when communication possibilities are added to a game to the theoretically larger set available if there is a reliable mediator available to collect information and recommend actions to the players.

Applications

Economic Applications

There is an enormous literature that uses signaling models in applications. Riley's [64] survey contains extended discussion of some of the most important applications. What follows is a brief discussion of some central ideas.

In a simple signaling game, one informed agent sends a single signal to one uninformed decision maker. This setting is reach enough to illustrate many important aspects of signaling, but it plainly limited. Interesting new issues arise if there are many informed agents, if there are many decision makers, and if the interaction is repeated. Several of the models below add some or all of these novel features to the basic model.

Advertising Advertisements are signals. Models similar to the standard model can explain situations in which higher levels of advertisement can lead consumers to believe the quality of the good is higher. In a separating equilibrium, advertising expenditures fully reveal quality. As in all costly signaling models, it is not important that there be a direct relationship between quality and signal, it is only necessary that firms with higher quality have lower marginal costs of advertising. Hence simply "burning money" or sending a signal that lowers utility by an amount independent of quality and response can be informative. The consumer may obtain full information in equilibrium, but someone must pay the cost of advertising. There are other situations where it is natural for the signal to be linked to the quality of the item. Models of verifiable information are appropriate in this case. When the assumptions of Proposition 5 hold, one would expect consumers to obtain all relevant information through disclosures without wasteful expenditures on signaling. Finally, cheap talk plays a role in some markets. One would expect costless communication to be informative in environments where heterogeneous consumers would like to identify the best product. Cheap talk can create more efficient matching of product to consumer. Here communication is free although in leading models separating equilibria do not exist.

Limit Pricing Signaling models offer one explanation for the phenomenon of limit pricing. An incumbent firm have private information about its cost. Potential entrants use the pricing behavior of the firm to draw inferences about the incumbent's cost, which determines profitability of entry. Milgrom and Roberts [54] construct an equilibrium in which the existence of incomplete information distorts prices: Relative to the full information model, the incumbent charges lower prices in order to signal that the market is relatively unprofitable. This behavior has the flavor of classical models of limit pricing, with one important qualification. In a separating equilibrium the entrant can infer the true cost of the incumbent and therefore the low price charged by the incumbent firm fails to change the entry decision.

Bargaining Several authors have proposed bargaining models with incomplete information to study the existence and duration of strikes [30,71]. If a firm with private information about its profitability makes a take-it-or-leave it offer to a union, then the strategic interaction is a simple signaling model in which the magnitude of the offer may serve as a signal of the firm's profitability. Firms with low profits are better able to make low wage offers to the union because the threat of a strike is less costly to a firm with low profits than one with high profits. Consequently settlement offers may reveal information. Natural extensions of this model permit counter offers. The variation of the model in which the uninformed agent makes offers and the uninformed agent accepts and rejects is formally almost identical to the canonical model of price discrimination by a durable-goods monopolist [4,39].

Finance Simple signaling arguments provide potential explanations for firms' choices of financial structure. Classic arguments due to Modigliani and Miller [58] and imply that firms' profitability should not depend on their choice of capital structure. Hence this theory cannot organize empirical regularities about firm's capital structure. The Modigliani–Miller theorem assumes that the firm's managers, shareholders, and potential shareholders all have access to the same information. An enormous literature assumes instead that the firm's managers have superior information and use corporate structure to signal profitability.

Leland and Pyle [47] assume that insiders are risk averse so they would prefer to diversify their personal holdings rather than maintain large investments in their firm. The value of diversification is greater the lower the quality of the firm. Hence when insiders have superior information than investors, there will be an incentive for the insiders of highly profitable firms to maintain inefficiently large investments in their firm in order to signal profitability to investors.

Dividends are taxed twice under the United States tax code, which raises the question of why firms would issue dividends when capital gains are taxed at a lower rate. A potential explanation for this behavior comes from a model in which investors have imperfect information about the future profitability of the firm and profitable firms are more able than less profitable firms to distribute profits in the form of dividends (see [14]).

Reputation Dynamic models of incomplete information create the opportunity for the receiver to draw inferences about the sender's private information while engaging in an extended interaction. Kreps and Wilson [44] and Milgrom and Roberts [55] provided the original treatments of reputation formation in games of incomplete information. Motivated by the limit pricing, their models examined the interaction of a single long-lived incumbent facing a sequence of potential entrants. The entrants lack information about the willingness of the incumbent to tolerate entry. Pricing decisions of the incumbent provide information to the entrants about the profitability of the market.

In these models, signals have implications for both current and future utility. The current cost is determined by the effect the signal has on current payoffs. In Kreps–Wilson and Milgrom–Roberts, this cost is the decrease in current profits associated with charging a low price. In other models (for example [59,70]) the actual signal is costless, but it has immediate payoff implications because of the response it induces. Signals also have implications for future utility because inferences about the sender's private information will influence the behavior of the opponents in future periods. Adding concern for reputation to a signaling game will influence behavior, but whether it leads to more or less informative signaling depends on the application.

Signaling in Biology

Signaling is important in biology. In independent and almost contemporaneous work, Zahavi [76] proposed a signaling model that shared the essential features of Spence [72]'s model of labor-market signaling. Zahavi observed that there are many examples in nature of animals apparently excessive physical displays. It takes energy to produce colorful plumage, large antlers, or loud cries. Having a large tail may actually make it harder for peacocks to flea predators. If a baby bird makes a loud sound to get his mother's attention, he may attract a dangerous predator. Zahavi argued that costly signals could play a role in sexual selection. In Zahavi's basic model, the sender is a male and the receiver is a female of the same species. Females who are able to mate with healthier males are more likely to have stronger children, but often the quality of a potential mate cannot be observed directly. Zahavi argued that if healthier males could produce visible displays more cheaply than less healthy males, then females would be induced to use the signals when deciding upon a mate. Displays may impose costs that "handicap" a signaler, but displays would persist when additional reproductive success compensates for their costs. Zahavi identifies a single-crossing condition as a necessary condition for the existence of costly signals.

The development of signaling in biology parallels that in economics, but there are important differences. Biology replaces the assumption of utility maximization and equilibrium with fitness maximization and evolutionary stability. That is, their models do not assume that animals consciously select their signal to maximize a payoff. Instead, the biological models assume that the process of natural selection will lead to strategy profiles in which mutant behavior has lower reproductive fitness than equilibrium behavior. This notion leads to static and dynamic solution concepts similar to Nash Equilibrium and its refinements. Fitness in biological models depends on contributions from both parents. Consequently, a full treatment of signaling must take into account population genetics. Grafen [34] discusses these issues and Grafen [33] and Siller [69] provide further theoretical development of the handicap theory. Finally, one must be careful in interpreting heterogeneous quality in biological models. Natural selection should operate to eliminate the least fit genes in a population. To the extent that this arises, there is pressure for quality variation within a population to decrease over time. The existence of unobserved quality variations needed for signaling may be the result of relatively small variations about a population norm.

While most of the literature on signaling in biology focuses on the use of costly signals, there are also situations in which cheap talk is effective. A leading example is the "Sir Philip Sidney Game," originally developed by John Maynard Smith [53] to illustrate the value of costly communication between a mother and child. The child has private information about its level of hunger and the mother must decide to feed the child or keep the food for itself. Since the players are related, survival of one positively influences the fitness of the other. This creates a common interest needed for cheap-talk communication. There are two ways to model communication in this environment. The first is to assume that signaling is costly, with hungrier babies better able to communicate their hunger. This could be because the sound of a hungry baby is hard for sated babies to imitate or it could be that crying for food increases the risk of predation and that this risk is relatively more dangerous to well fed chicks than to starving ones (because the starving chicks have nothing to lose). This game has multiple equilibria in which signals fully reveal the state of the baby over a range of values (see [46,53]). These papers look at a model in which both mother and child have private information. Alternatively, Bergstrom and Lachmann [13] study a cheap-talk version of the game. Here there may be an equilibrium outcome in which the baby bird credibly signals whether or not he is hungry. Those who signal hunger get fed. The others do not. Well fed baby birds may wish to signal that they are not hungry in order to permit the mother to keep food for herself. Such an equilibrium exists if the fraction of genes that mother and child share is large and the baby is already well fed.

Political Science

Signaling games have played an important role in formal models of political science. Banks [7] reviews models of agenda control, political rhetoric, voting, and electoral competition. Several important models in this area are formally interesting because they violate the standard assumptions frequently satisfied in economic models. I describe two such models in this subsection.

Banks [6] studies a model of agenda setting in which the informed sender proposes a policy to a receiver (decision-maker), who can either accept or reject the proposal. If the proposal is accepted, it becomes the outcome. If not, then the outcome is a fall-back policy. The fall-back policy is known only to the sender. In this environment, the sender's strategy may convey information to the decision maker. Signaling is costly, but, because the receiver's set of actions in binary, fully revealing equilibria need not exist. Refinements limit the set of predictions in this model to a class of outcomes in which only one proposal is accepted in equilibrium (and that this proposal is accepted with probability one), but there are typically a continuum of possible equilibrium outcomes.

Matthews [52] develops a cheap-talk model of veto threats. There are two players, a Chooser (C), who plays the role of receiver, and a Proposer (P), who plays the role of sender. The players have preferences that are represented by single-peaked utility functions which depend on the real-valued outcome of the game and an ideal point. P's ideal point is common knowledge. C's ideal point is her private information, drawn from a prior distribution that has a smooth positive density on a compact interval, $[\underline{t}, \bar{t}]$. The game form is simple: C learns her type, then sends a cheap-talk signal to P, who responds with a proposal. C then either accepts or rejects the proposal. Accepted proposals become the outcome of the game. If C rejects the proposal, then the outcome is the status quo point.

As usual in cheap-talk games, this game has a babbling outcome in which C's message contains no information and P makes a single, take-it-or-leave-it offer that is accepted with probability strictly between 0 and 1. Matthews shows there may be equilibria in which two outcomes are induced with positive probability (size-two equilibria), but size $n > 2$ (perfect Bayesian) equilibria never exist. In a size-two equilibrium, P offers his ideal outcome to those types of C whose message indicates that their ideal point is low; this offer is always accepted in equilibrium. If C indicated that his ideal point is high, P makes a compromise offer that is sometimes accepted and sometimes rejected.

Future Directions

The most exciting developments in signaling games in the future are likely to come from interaction between economics and other disciplines.

Over the last ten years the influence of behavioral economists have led the profession to rethink many of its fundamental models. An explosion of experimental studies have already influenced the interpretation of signaling models and have led to a re-examination of basic assumptions. There is evidence that economic actors lack the strategic sophistication assumed in equilibrium models. Further, economic agents may be motivated by more than their material well being. Existing experimental evidence provides broad support for many of the qualitative predictions of the theory (Banks, Camerer, and Porter [5] and Brandts and Holt [17]), but also suggests ways in which the theory may be inadequate.

The driving assumption of signaling models is that when informational asymmetries exist, senders will attempt to lie for strategic advantage and that sophisticated receivers will discount statements. These assumptions may be reconsidered in light of experimental evidence that some agents will behave honestly in spite of strategic incentives to lie. For example, Gneezy [32] and Hurkens and Kartik [40] present experimental evidence that some agents are reluctant to lie even when there is a financial gain from doing so. There is evidence from other disciplines that some agents are unwilling or unable to manipulate information for strategic advantage and that people may be well equipped to detect these manipulations in ways that are not captured in standard models (see, for example, Ekman [27] or Trivers [73]). Experimental evidence and, possibly, results from neuroscience may demonstrate that the standard assumption that some agents cannot manipulate information for their strategic advantage (or that other agents have ability to see through deception) will inform the development of novel models of communication that include behavioral types. Several papers study the implications of including behavioral types into the standard paradigm. The reputation models of Kreps and Wilson [44] and Milgrom and Roberts [54] are two early examples. Papers on communication by Chen [19], Crawford [23], Kartik [41], and Olszewski [61] are more recent examples. New developments in behavioral economics will inform future theoretical studies.

There is substantial interest in signaling in philosophy. Indeed, the philosopher David Lewis [48] (first published in 1969) introduced signaling games prior to the contributions of Spence and Zahavi. Recently linguists have been paying more attention to game-theoretic ideas. Benz, Jäger and Van Rooij [12] collects recent work that attempts to formalize ideas from linguistic philosophy due to Grice [36]. While there have been a small number of contributions by economists in this area (Rubinstein [66] and Sally [67] are examples), there is likely to be more active interaction in the future.

Finally, future work may connect strategic aspects of communication to the actual structure of language. Blume [15], Cucker, Smale, Zhou [25], and Nowak and Krakauer [60] present dramatically different models on how structured communication may result from learning processes. Synthesizing these approaches may lead to fundamental insights on how the ability to send and receive signals develops.

Acknowledgments

I thank the Guggenheim Foundation, NSF, and the Secretaría de Estado de Universidades e Investigación del Ministerio de Educación y Ciencia (Spain) for financial support and Jose Penalva for his comments. I am grateful to the Departament d'Economia i d'Història Econòmica and Institut d'Anàlisi Econòmica of the Universitat Autònoma de Barcelona for hospitality and administrative support.

Bibliography

Primary Literature

1. Ambrus A, Takahashi S (2008) Multi-sender cheap talk with restricted state space. Theor Econ 3(1):1–27
2. Aumann R, Hart S (2003) Long cheap talk. Econometrica 71(6):1619–1660
3. Aumann RJ (1990) Nash equilibria and not self-enforcing. In: Gabszewicz JJ, Richard J-F, Wolsey LA (eds) Economic Decision Making: Games, Econometrics and Optimisation. Elsevier, Amsterdam, pp 201–206
4. Ausubel LM, Deneckere RJ (1989) Reputation in bargaining and durable goods monopoly. Econometrica 57(3):511–531
5. Banks J, Camerer C, Porter D (1994) An experimental analysis of nash refinements in signaling games. Games Econ Behav 6(1):1–31
6. Banks JS (1990) Monopoly agenda control and asymmetric information. Q J Econ 105(2):445–464
7. Banks JS (1991) Signaling Games in Political Science. Routledge, Langhorne
8. Banks JS, Sobel J (1987) Equilibrium selection in signaling games. Econometrica 55(3):647–661
9. Battaglini M (2002) Multiple referrals and multidimensional cheap talk. Econometrica 70(4):1379–1401
10. Ben-Porath E (2003) Cheap talk in games with incomplete information. J Econ Theory 108:45–71
11. Ben-Porath E, Dekel E (1992) Signaling future actions and the potential for sacrifice. J Econ Theory 57:36–51
12. Benz A, Jäger G, Van Rooij R (eds) (2005) Game Theory and Pragmatics. Palgrave MacMillan, Houndmills, Basingstoke
13. Bergstrom CT, Lachmann M (1998) Signalling among relatives. III. talk is cheap. Proc Natl Acad Sci USA 95:5100–5015
14. Bhattacharya S (1979) Imperfect information, dividend policy, and the bird in the hand fallacy. Bell J Econ 10(1):259–270
15. Blume A (2000) Coordination and learning with a partial language. J Econ Theory 95:1–36
16. Blume A, Board O, Kawamura K (2007) Noisy talk. Theor Econ 2(4):396–440
17. Brandts J, Holt CA (1992) An experimental test of equilibrium dominance in signaling games. Am Econ Rev 82(5):1350–1365
18. Chakraborty A, Harbaugh R (2007) Comparative cheap talk. J Econ Theory 132(1):70–94
19. Chen Y (2005) Perturbed communication games with honest senders and naive receivers. Technical report. Arizona State University, Tempe
20. Chen Y, Kartik N, Sobel J (2007) On the robustness of informative cheap talk. Technical report, UCSD. In: Econometrica 76(1):117–136, Blackwell Synergy
21. Cho I-K, Kreps DM (1987) Signaling games and stable equilibria. Q J Econ 102(2):179–221
22. Cho I-K, Sobel J (1990) Strategic stability and uniqueness in signaling games. J Econ Theory 50(2):381–413
23. Crawford VP (2003) Lying for strategic advantage: Rational and boundedly rational misrepresentation of intentions. Am Econ Rev 93(1):133–149
24. Crawford VP, Sobel J (1982) Strategic information transmission. Econometrica 50(6):1431–1451
25. Cucker F, Smale S, Zhou D-X (2004) Modeling language evolution. Found Comput Math 4(3):315–343
26. Edlin AS, Shannon C (1998) Strict single-crossing and the strict spence-mirrlees condition: A comment on monotone comparative statics. Econometrica 66(6):1417–1425
27. Ekman P (2001) Telling Lies. W.W. Norton, New York
28. Farrell J (1993) Meaning and credibility in cheap-talk games. Games Econ Behav 5(4):514–531
29. Forges F (1990) Equilibria with communication in a job market example. Q J Econ 105(2):375–398
30. Fudenberg D, Tirole J (1983) Sequential bargaining with incomplete information. Rev Econ Stud 50(2):221–247
31. Giovannoni F, Seidmann DJ (2007) Secrecy, two-sided bias and the value of evidence. Games Econ Behav 59(2):296–315
32. Gneezy U (2005) Deception: The role of consequences. Am Econ Rev 95(1):384–394
33. Grafen A (1990) Biological signals as handicaps. J Theor Biol 144:517–546
34. Grafen A (1990) Sexual selection unhandicapped by the fisher process. J Theor Biol 144:473–516
35. Green JR, Stokey NL (2007) A two-person game of information transmission. J Econ Theory 135(1):90–104
36. Grice HP (1991) Studies in the Way of Words. Harvard University Press, Cambridge
37. Grossman S (1981) The role of warranties and private disclosure about product quality. J Law Econ 24:461–483
38. Grossman S, Perry M (1987) Perfect sequential equilibria. J Econ Theory 39:97–119
39. Gul F, Sonnenchein H, Wilson R (1986) Foundations of dynamic monopoly and the coase conjecture. J Econ Theory 39:155–190
40. Hurkens S, Kartik N (2006) (When) Would I lie to you? Comment on Deception: The role of consequences. Technical report, UCSD
41. Kartik N (2005) Information transmission with almost-cheap talk. Technical report, UCSD
42. Kohlberg E, Mertens J-F (1986) On the strategic stability of equilibria. Econometrica 54(5):1003–1037
43. Kreps DM, Sobel J (1994) Signalling. In: Aumann RJ, Hart S (eds) Handbook of game theory: with economics applications. Handbooks in Economics, no 11, vol 2. Elsevier, Amsterdam, chap 25, pp 849–868
44. Kreps DM, Wilson R (1982) Reputation and imperfect information. J Econ Theory 27:253–277
45. Krishna V, Morgan J (2001) A model of expertise. Q J Econ 116(2):747–775
46. Lachmann M, Bergstrom CT (1998) Signalling among relatives. II. beyond the tower of babel. Theor Popul Biol 54:146–160
47. Leland HE, Pyle DH (1977) Informational asymmetries, financial structure, and financial intermediation informational asymmetries, financial structure, and financial intermediation. J Finance 32(2):371–387

48. Lewis D (2002) Convention: A Philosophical Study. Blackwell, Oxford
49. Mailath GJ (1987) Incentive compatibility in signaling games with a continuum of types. Econometrica 55(6):1349–1365
50. Mailath GJ, Okuno-Fujiwara M, Postlewaite A (1993) On belief based refinements in signaling games. J Econ Theory 60(2):241–276
51. Mathews SA, Okuno-Fujiwara M, Postlewaite A (1991) Refining cheap-talk equilibria. J Econ Theory 55(2):247–273
52. Matthews SA (1989) Veto threats: Rhetoric in a bargaining game. Q J Econ 104(2):347–369
53. Maynard Smith J (1991) Honest signalling: the Philip Sidney game. Anim Behav 42:1034–1035
54. Milgrom P, Roberts J (1982) Limit pricing and entry under incomplete information: An equilibrium analysis. Econometrica 50(2):443–459
55. Milgrom P, Roberts J (1982) Predation, reputation, and entry deterence. J Econ Theory 27:280–312
56. Milgrom PR (1981) Good news and bad news: Representation theorems. Bell J Econ 21:380–391
57. Milgrom PR, Shannon C (1994) Monotone comparative statics. Econometrica 62(1):157–180
58. Modigliani F, Miller MH (1958) The cost of capital, corporation finance and the theory of investment. Am Econ Rev 48(3):261–297
59. Morris S (2001) Political correctness. J Political Econ 109:231–265
60. Nowak MA, Krakauer DC (1999) The evolution of language. Proc Natl Acad Sci 96(14):8028–8033
61. Olszewski W (2004) Informal communication. J Econ Theory 117:180–200
62. Rabin M (1990) Communication between rational agents. J Econ Theory 51:144–170
63. Rabin M, Farrell J (1996) Cheap talk. J Econ Perspectives 10(3):103–118
64. Riley JG (2001) Silver signals: Twenty-five years of screening and signaling. J Econ Lit 39(2):432–478
65. Rothschild M, Stiglitz J (1976) Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. Q J Econ 90(4):629–649
66. Rubinstein A (2000) Economics and language. Cambridge University Press, New York
67. Sally D (2005) Can I say, bobobo and mean, there's no such thing as cheap talk? J Econ Behav Organ 57(3):245–266
68. Seidmann DJ, Winter E (1997) Strategic information transmission with verifiable messages. Econometrica 65(1):163–169
69. Siller S (1998) A note on errors in grafen's strategic handicap models. J Theor Biol 195:413–417
70. Sobel J (1985) A theory of credibility. Rev Econ Stud 52(4):557–573
71. Sobel J, Takahashi I (1983) A multistage model of bargaining. Rev Econ Stud 50(3):411–426
72. Spence AM (1974) Market Signaling. Harvard University Press, Cambridge
73. Trivers RL (1971) The evolution of reciprocal altruism. Q Rev Biol 46:35–58
74. Vida P (2006) Long Pre-Play Communication in Games. Ph D thesis, Autonomous University of Barcelona
75. Wilson C (1977) A model of insurance markets with incomplete information. J Econ Theory 16:167–207
76. Zahavi A (1975) Mate selection- a selection for a handicap. J Theor Biol 53:205–214

Books and Reviews

Admati A, Perry M (1987) Strategic delay in bargaining. Rev Econ Stud 54:345–364
Austen-Smith D (1990) Information transmission in debate. Am J Political Sci 34(1):124–152
Battigalli P, Siniscalchi M (2002) Strong belief and forward induction reasoning. J Econ Theory 106(2):356–391
Bernheim BD (1994) A theory of conformity. J Political Econ 102(5):841–877
Blume A, Kim Y-G, Sobel J (1993) Evolutionary stability in games of communication. Games Econ Behav 5:547–575
Fudenberg D, Tirole J (1991) Game Theory. MIT Press, Cambridge
Gibbons R (1992) Game Theory for Applied Economists. Princeton University Press, Princeton
Kartik N, Ottaviani M, Squintani F (2007) Credulity, lies, and costly talk. J Econ Theory 134(1):93–116
Krishna V, Morgan J (2001) A model of expertise. Q J Econ 116(2):747–775
Lo P-Y (2006) Common knowledge of language and iterative admissibility in a sender-receiver game. Technical report. Brown University
Manelli AM (1996) Cheap talk and sequential equilibria in signaling games. Econometrica 64(4):917–942
Milgrom P, Roberts J (1986) Price and advertising signals of product quality. J Political Econ 94(4):796–821
Noldeke G, Samuelson L (1997) A dynamic model of equilibrium selection in signaling markets. J Econ Theory 73(1):118–156
Noldeke G, Van Damme E (1990) Signalling in a dynamic labour market. Rev Econ Stud 57(1):1–23
Ottaviani M, Sørensen PN (2006) Professional advice. J Econ Theory 126(1):120–142
Rabin M (1994) A model of pre-game communication. J Econ Theory 63(2):370–391
Ramey G (1996) D1 signaling equilibria with multiple signals and a continuum of types. J Econ Theory 69(2):508–531
Rasmusen EB (2006) Games and Information: An Introduction to Game Theory, 4th edn. Blackwell, New York
Riley JG (1979) Informational equilibrium. Econometrica 47:331–359
Sobel J, Stole L, Zapater I (1990) Fixed-equilibrium rationalizability in signaling games. J Econ Theory 52(2):304–331
Spence AM (1973) Job market signaling. Q J Econ 90:225–243
Swinkels JM (1999) Education signallling with preemptive offers. Rev Econ Stud 66(4):949–970

Single Column Modeling of Atmospheric Boundary Layers and the Complex Interactions with the Land Surface

Albert A. M. Holtslag, Gert-Jan Steeneveld
Department of Meteorology and Air Quality,
Wageningen University, Wageningen, The Netherlands

Article Outline

Glossary

Atmospheric boundary layer The Atmospheric Boundary Layer (ABL) is the lower part of the atmosphere which is directly influenced by the presence of the earth's surface. As such its major characteristics are turbulence and the diurnal cycle.

Diurnal cycle The depth of the dry atmospheric boundary layer (ABL) can vary over land between tens of meters during night up to kilometers during daytime (see Fig. 2). Over sea the depth is often typical a few hundred meters and rather constant on the time scale of a day.

Turbulence Turbulence in the atmospheric boundary layer is the three-dimensional, chaotic flow of air with time scales typically between a second and an hour. The corresponding length scales are from a millimeter up to the depth of the boundary layer (or more in the case of clouds). Turbulence in the ABL originates due to friction of the flow and heating (convection) at the surface.

Definition of the Subject

In this article we deal with the single column modeling of the Atmospheric Boundary layer (ABL) and the the complex interactions which may occur with the land surface. As such we review the major characteristics of the ABL over land, and summarize the basic parameterizations for the represesentation of atmospheric turbulence and the surface fluxes. The modeling principles are illustrated with the outcome of single-column models for a variety of conditions using field data and fine-scale model results. Our emphasis is on stable conditions which occur over land at night-time under clear skies. For readers not familiar with atmospheric turbulence and meteorological definitions, some background and basic definitions are also given.

Introduction

The Atmospheric Boundary Layer (ABL) is generally characterized by turbulence. Because of its capability to mix air with different properties efficiently, the representation of turbulence is directly relevant for atmospheric and environmental modeling. For instance, turbulence directly impacts on the transfer of momentum, sensible heat, water vapor, ozone, and methane, among many other quantities, between the earth's surface and the atmosphere. Turbulence also defines the mixing of properties inside the atmospheric boundary layer, the transfer of quantities between the boundary layer and the clear or cloudy atmosphere aloft, and the mixing inside clouds.

Turbulence in the ABL is mainly due to the mechanical turbulence by vertical wind shear and turbulence by convection. Most of the atmosphere above the ABL is not turbulent, although turbulence can occur throughout the whole atmosphere. For instance, cumulus-type clouds, which may grow into thunderstorms, are always turbulent through convection produced by the heat released due to the condensation of water vapor. Turbulence can also occur in clear air above the ABL; most of this is produced in layers of strong vertical wind shear at the boundary between air masses (so-called 'Clear-Air Turbulence').

Because of the mixing capacity of turbulence, modeling atmospheric boundary layers is also relevant for many practical applications. For instance, chimney plumes are diluted and spread over larger volumes than they would be without turbulence. As such, strong local peaks of pollution are prevented and otherwise clean air is polluted. In practice turbulence may also cause engineering problems, because it shakes structures such as bridges, towers, and airplanes, causing failure of such systems in extreme cases. Turbulent fluctuations in the horizontal motions during severe storms can be fatal to tall buildings or bridges, particularly if resonance (e. g., forcing of a system at its natural frequency) occurs.

The correct formulation of the overall effects by turbulence, either inside or outside the atmospheric boundary layer, is an essential part of atmospheric models dealing with the prediction and study of weather, climate and air quality. These models are based on solving the equations dealing with atmosphere behavior. With state-of-the-art computers, the number of grid points in atmospheric models is limited to a number of typically 10^8 or so. This implies that on the regional and global scale the atmospheric model equations are usually applied too fairly large ‘air boxes’. Such boxes are often in the order of ten to hundred kilometers wide and ten to a few hundred meters thick. In these large boxes, smaller scale motions make air parcels interact and mix. For example, if a hot parcel is located next to a cold parcel, turbulent motion at their boundaries will heat the cool and cool the hot parcel. Thus, a closure formulation is needed to reproduce mixing by the turbulent motions into the model-resolved scales using the equations for the larger-scale ‘mean’ motions. It is important to realize that the closure formulation needs to be expressed in terms of variables available in the modeling context. This is called a ‘parameterization’.

In this contribution we provide an overview of the modeling principles, the turbulent closures and parameterizations in use for of the atmospheric boundary layer, where we emphasize the modeling and parameterization of turbulence in the atmospheric boundary layer without clouds. Additionally we discuss the performance of models in current use [7,28], and we study the impact of the surface boundary condition over land [18].

Background

Atmospheric models for the forecasting and study of weather, climate, and air quality are typically based on integration of the basic equations governing atmospheric behavior. These equations are the gas law, the equation of continuity (mass), the first law of thermodynamics (heat), the conservation equations for momentum (the so-called “Navier–Stokes equations”), and usually equations expressing the conservation of moisture, trace gases and air pollutants. At one extreme, atmospheric models may deal with the world’s climate and climate change; at the other, they may account for the behavior of local flows at coasts, in mountain-valley areas, or even deal with individual clouds. This all depends on the selected horizontal modeling domain and the available computing resources.

Since there is an enormous range of scales in atmospheric motion and turbulence, there is a need to separate the scales of atmospheric turbulence from larger-scale motions. Let C denote an atmospheric variable, such as specific humidity. Then $\overline{C}$ represents a mean or “smoothed” value of C, typically taken on a horizontal scale of order 10 (or more) km and a corresponding time scale in the order of 10 min to one hour. A local or instantaneous value of C would differ from $\overline{C}$. Thus, we have

$$C = \overline{C} + c\,. \tag{1}$$

Here c represents the smaller-scale fluctuations. Note that we use lower case for the latter (often primes are used as well to indicate fluctuations). In principle, the fluctuations around the mean motion also reflect gravity waves and other smaller scale motions, in addition to turbulence. Gravity waves often co-exist with turbulence or are generated by turbulence. If the wind at the same time is weak, there may be no turbulence at all. Anyhow, if turbulence exists, it is usually more important for most atmospheric applications, because it mixes more efficiently than the other small-scale motions.

To make the mathematical handling of c tractable, it must satisfy the so-called “Reynolds postulates”. These require, for example, that $\overline{c} = 0$ and that small- and larger-scale values must not be correlated. After a quantity has been averaged to create a larger-scale quantity, further averaging should produce no further changes, in order for this postulate to apply. The mean of the summation of two variables A and C will produce $\overline{A \pm C} = \overline{A} \pm \overline{C}$. A further condition is that a mean variable $\overline{C}$ must be differentiable, since differentials show up in the atmospheric equations (see below). In practice, not all these conditions are rigorously satisfied. If the Reynolds postulates are fulfilled, then the averaging for the product of two variables provides

$$\overline{AC} = \overline{A}\,\overline{C} + \overline{ac}\,. \tag{2}$$

The second term at the right hand side of Eq. (2) is known as the turbulent covariance. Similarly, the turbulence variance of a quantity is given by $\overline{C^2} - (\overline{C})^2$ (which is the square of the standard deviation).

If in Eq. (2), the variable A represents one of the velocity components (U, V, W in the x, y, z direction, respectively), then $\overline{AC}$ is the total flux of C and the second term at the right hand side of Eq. (2) represents a turbulent flux of C. For instance, $\overline{uc}$ and $\overline{wc}$ are the horizontal and vertical turbulent fluxes of the variable C, respectively. Here u and w are the turbulent fluctuations of the horizontal and vertical velocities. Near the surface, the mean vertical wind $\overline{W}$ is usually small, and thus the total vertical fluxes are normally dominated by the turbulent contributions.

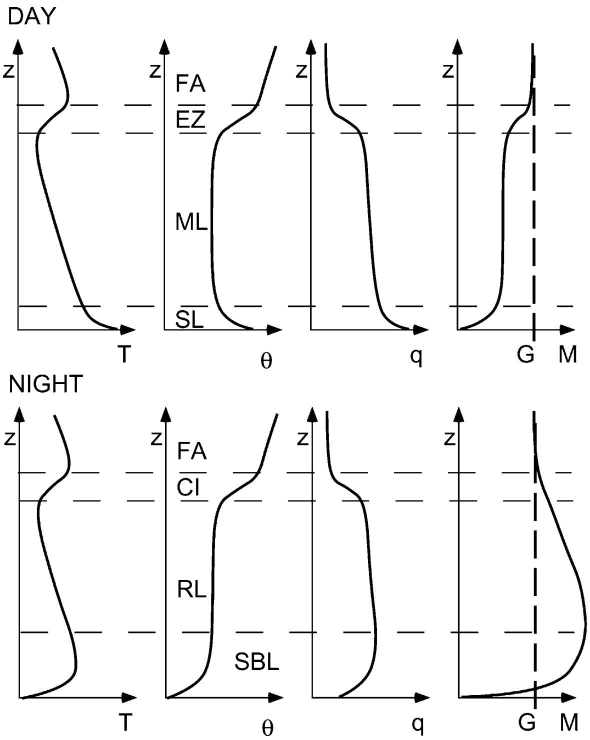
Atmospheric Boundary-Layer Structure

Turbulent fluctuations, variances and fluxes of variables are influenced by the vertical boundary-layer structure. Here the variation of temperature in the atmospheric boundary-layer plays an important role. Since pressure decreases with altitude, air parcels, which are forced to rise (sink), do expand (compress). According to the first law of thermodynamics, a rising (sinking) parcel will cool (warm) if there is no additional energy source such as condensation of water vapor. Then this is called a dry adiabatic process.

It can be shown that in the atmospheric boundary layer, the temperature (T) variation with height for a dry adiabatic process is $\mathrm{d}T/\mathrm{d}z = -g/C_\mathrm{p}$ (here g is gravity constant and C_p is specific heat at constant pressure). The value for g/C_p is approximately 1 K per 100 m. An atmospheric layer which has such a temperature variation with height, is called neutral for dry air (at least when there is no convection arising from other levels). In that case $\Theta = T + (g/C_\mathrm{p})z$ is constant, where Θ is called the potential temperature (Note that the previous definition for potential temperature is not accurate above the boundary layer). Since air normally contains water vapor and because moist air is lighter than dry air, we have to correct for the influence of this on vertical motions. Consequently, a virtual potential temperature is defined as $\Theta_\mathrm{v} = \Theta(1 + 0.61q)$, where q is the specific humidity (defined as the mass of water vapor per unit mass of moist air).

In a neutral layer with constant Θ_v, vertical motions of moist (not saturated) air can maintain themselves. If the virtual potential temperature of the atmospheric layer increases with height, vertical displacements are suppressed. This is called a stable condition (or 'inversion'). At the other hand, when the virtual potential temperature decreases with height, vertical fluctuations may be accelerated. Consequently this is called an unstable condition. Thus in considerations with turbulent fluctuations and atmospheric stability, we have to deal with the virtual potential temperature and not with the actual temperature. Similarly, the vertical flux of sensible heat is connected to turbulent fluctuations of (virtual) potential temperature; e. g. it reads as $\overline{w\theta_\mathrm{v}}$ (in m K/s). The latter relates directly to the energy per time and unit area H by $H = \rho C_\mathrm{p}\overline{w\theta_\mathrm{v}}$ (in W/m^2), where ρ is density of the air (in kg/m^3).

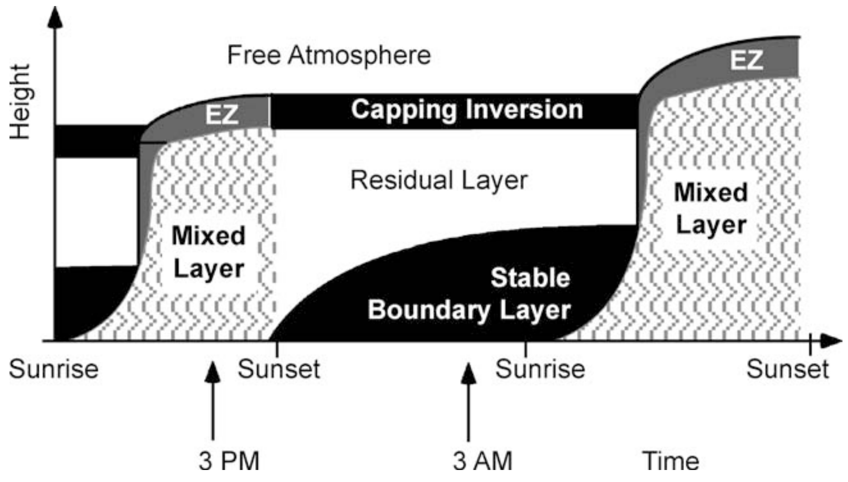
Figure 1 (after [30]), provides the typical, idealized, mean vertical profiles for temperature T, potential temperature Θ, specific humidity q, in addition to the horizontal wind M (defined by $M^2 = U^2 + V^2$). These profiles apply for an atmospheric boundary layer over land in clear sky conditions in the afternoon and around midnight. Note that in the free atmosphere the horizontal wind is mostly a result of the acting of the larger scale pressure differences and the Coriolis force due to the rotation of the earth (but other effects may play a role as well). The resulting wind is known as the 'geostrophic' wind and indicated with G in Fig. 1 (see dashed line). In the daytime boundary layer the actual wind is smaller due to surface friction, while at clear nights the actual wind away from the surface may be substantially stronger than G due to inertial effects (resulting in the so-called 'low level jet').



Single Column Modeling of Atmospheric Boundary Layers and the Complex Interactions with the Land Surface, Figure 1
Idealized vertical profiles of mean variables in the Atmospheric Boundary Layer over land in fair weather (after [30]). See text for additional information

The temporal variation of the mean boundary-layer profiles over land can be quite substantial due to the strong diurnal variation of solar incoming radiation and the nighttime cooling at the land surface. During daytime the turbulent boundary layer may grow to several kilometers into the non-turbulent 'free atmosphere' (indicted as FA in Fig. 1). At night the turbulent part of the stable boundary layer (SBL) may only extend up to a few hundred meters or less (the lowest dashed line in the lower figure). An idealized picture for the temporal variation of the boundary layer over land is given in Fig. 2 (after [30]). Here the arrows with local time indications refer to the day and nighttime figures of Fig. 1.

Figure 1 also indicates that the boundary layer during daytime shows a three-layer structure: an unstable 'surface layer (SL)', a 'well-mixed layer (ML)' with rather uniform (virtual) potential temperature, and a stably-stratified 'entrainment zone (EZ)'. In the latter zone, turbulence acts to exchange heat, momentum, water-vapor and trace gasses



Single Column Modeling of Atmospheric Boundary Layers and the Complex Interactions with the Land Surface, Figure 2
Idealized diurnal evolution of the Atmospheric Boundary Layer over land in fair weather (after [30])

between the boundary layer and the free atmosphere. During nighttime, often the vertical structure of the previous day persists above the SBL. As such a ‘residual layer (RL)’ with sporadic turbulence (remaining from the previous day) can be identified as well as a ‘capping inversion (CI)’.

Modeling Basics

The challenge of modeling the atmospheric boundary layer is the prediction of the temporal variation of the vertical and horizontal structures in response to the influence of the major processes acting in the atmosphere and at the earth’s surface. As such the governing equations have to be integrated. In practice, the variables are split into ‘mean’ larger-scale motions and smaller-scale fluctuations as in Eq. (1). Inserting this into the basic equations and after averaging this provides a set of equations for the behavior of the larger-scale (mean) variables. The larger-scale variables are then used explicitly in atmospheric models. This can be demonstrated as follows below.

The general character of any of the budget equations dealing with atmospheric motions is

$$\frac{\mathrm{D}C}{\mathrm{D}t} = S_i \,. \tag{3a}$$

Here S_i represents the subsequent sources and sinks for the variable C (such as radiation or chemistry effects). The notation $\mathrm{D}C/\mathrm{D}t$ represents the total rate of change for the variable C by local changes ($\partial/\partial t$), and changes transported with the fluid motion in the three directions. As such, we have

$$\frac{\partial C}{\partial t} + U\frac{\partial C}{\partial x} + V\frac{\partial C}{\partial y} + W\frac{\partial C}{\partial z} = S_i \,. \tag{3b}$$

Here U, V, W are the wind speeds in the three directions x, y, z, respectively.

If in the atmospheric motion each variable is split into a mean component and a fluctuation then (3b) provides after Reynolds-averaging, some algebraic manipulations and simplifying assumptions, a budget equation for the mean variable $\overline{C}$. This reads as

$$\begin{aligned}\frac{\mathrm{D}\overline{C}}{\mathrm{D}t} &= \frac{\partial\overline{C}}{\partial t} + \overline{U}\frac{\partial\overline{C}}{\partial x} + \overline{V}\frac{\partial\overline{C}}{\partial y} + \overline{W}\frac{\partial\overline{C}}{\partial z}\\ &= \overline{S_i} - \frac{\partial\overline{uc}}{\partial x} - \frac{\partial\overline{vc}}{\partial y} - \frac{\partial\overline{wc}}{\partial z}\,.\end{aligned} \tag{4}$$

We may note that in the derivation of (4), single terms representing fluctuations have disappeared (as above in Eq. (2)). However, terms involving the product of two fluctuations did remain.

Thus because the basic equations are nonlinear, the budget equations for the mean variables contain terms involving smaller-scale motions. The latter terms are of the form of a divergence of fluxes produced by such motions in the three directions and appear as the last three terms in Eq. (4). These motions are said to be sub-grid and consequently, closure formulations or parameterizations are needed to introduce mixing by the smaller-scale, sub-grid, motions into the equations for the larger-scale motions (as resolved by the model). Note that additional terms may also appear in (4) when the source or sink term S_i incorporates nonlinear effects (such as in the case of chemistry).

The atmospheric model equations can also be applied on much smaller spatial and temporal scales then discussed here, for instance by using vertical and horizontal grid elements of 10 to 100 m, and time steps of seconds only. It is important to realize that in such cases a significant part of the turbulent fluctuations are resolved by the model equations. This type of modeling is nowadays known as 'Large-eddy simulation (LES)'. This has become a powerful and popular tool in the last decade to study turbulence in clear and cloudy boundary layers under well-defined conditions. It is important to realize that in the case of LES the simplifying assumptions leading to Eq. (2) are normally not valid.

A special and simple form of Eq. (4) arises for horizontally homogeneous conditions. In such cases the terms including horizontal derivatives are negligible. If in addition the mean vertical wind is small and if there are no other sources and sinks, then (4) provides

$$\frac{\partial \overline{C}}{\partial t} = -\frac{\partial \overline{wc}}{\partial z} \, . \tag{5}$$

This equation is known as the one-dimensional, vertical diffusion equation. It shows that the local time rate of change for the mean of a variable (such as temperature or wind) at a certain height, is given by the divergence of the turbulent (corresponding heat or momentum) flux in the vertical direction. As such, information on the turbulent flux may produce a local forecast of the variation of a mean variable (but only under the simplifications mentioned).

Equation (5) can be seen as the basis of a single-column model where only local information of the atmosphere is relevant. However, normally the other terms in (4) are also relevant, in particular the terms with mean wind speed (the so-called "advection terms"). This means that in general the budget equations for momentum, heat, and the various scalars are closely coupled in any atmospheric model. Still one can solve for the local time rate of change in a single column model once the advection terms are known form observations or other means. This approach is widely adopted to study atmospheric boundary layers in comparison with observations on a local scale.

Before we proceed with more detailed parameterizations for the fluxes in the boundary layer, let us deal with the derivation of the surface fluxes. These fluxes enter as boundary conditions when solving the budget equations for all the relevant mean variables (in any approach). It is important to realize that near the surface, the average wind must vanish because the mean wind is zero at the earth's surface. At the other hand, we know from observations that the fluxes of heat, momentum and trace gasses are nonzero. Consequently, it is convenient to model an 'effective' surface flux $\overline{wc_0}$ of a conserved variable due to the combined effect of molecular diffusion and turbulence at the surface. This can be achieved by writing

$$\overline{wc_0} = \beta_t w_t (C_0 - C_a) \, . \tag{6}$$

Here C_0, and C_a are the values of the transported variable at the surface and in the air, respectively; β_t is a transfer coefficient, and w_t is an effective transport velocity representing the turbulence. For example, in near-neutral conditions the effective transport velocity is well represented by the well-known surface friction velocity u_{*0}. Then it can be shown that $\beta_t = \kappa / \ln(z/z_0)$, where κ is the 'Von Karman' constant (often specified as $\kappa \cong 0.4$), z is the corresponding height of C_a in the lowest part of the boundary layer and z_0 is the so-called surface roughness length for the variable C. We refer to the literature for a more detailed treatment (e. g., Beljaars and Holtslag [4]).

Local Mixing Parameterization

To solve the budget Eq. (4) for all the mean atmospheric variables involved, the terms involving turbulent fluxes need to be parametrized. As mentioned before, this means that the fluxes need to be expressed in terms of available mean model quantities, both in the atmosphere and at the surface. Once this has been achieved, the atmospheric model equations can be integrated. Thus, starting with proper initial values, new values can be calculated for the following time step and so on.

The most frequently used parameterization for environmental and atmospheric models, is known as first-order closure or often also called K-theory. In this theory it is assumed that the flux $\overline{wc}$ of a variable C in the vertical direction z, is down the vertical gradient of the mean concentration of C per unit mass. Thus

$$\overline{wc} = -K_c \frac{\partial \overline{C}}{\partial z} \, . \tag{7}$$

Here, K_c is known as the 'eddy-diffusivity' or mixing coefficient for the variable C. Similarly, the horizontal fluxes can be represented in terms of horizontal gradients. Note that the corresponding eddy-diffusivities typically are not constant, but that they generally depend on properties of the flow and the variable of interest. This also means that normally no analytic solutions are possible, not even for the simple case in which Eqs. (5) and (8) are combined.

We may note that the dimension of an eddy-diffusivity is a length scale ℓ times a velocity scale. These are proportional to the products of effective eddy sizes and eddy velocities in the corresponding directions. Often a diagnostic expression is used for the eddy-diffusivity, on basis of what is called 'mixing length theory' (in analogy with molecular diffusion). The result reads as

$$K_c = \ell^2 S f(\mathrm{Ri})\,. \tag{8}$$

Here S is vertical wind shear (that is the variation of mean horizontal wind with height). Note that the combination ℓS in Eq. (9) has units of velocity. In Eq. (8), $f(\mathrm{Ri})$ denotes a functional dependence on local stability as represented by the gradient Richardson-number Ri defined by

$$\mathrm{Ri} = \frac{g}{\overline{\Theta_\mathrm{v}}} \frac{\partial\overline{\Theta_\mathrm{v}}/\partial z}{(\partial\overline{U}/\partial z)^2 + (\partial\overline{V}/\partial z)^2}\,. \tag{9}$$

Here g is the acceleration due to gravity, and $\overline{\Theta_\mathrm{v}}$ is the mean 'virtual potential temperature'.

The specification of the length scale ℓ is not at all straightforward, except near the surface where so-called 'surface-layer similarity theory' (see cited literature) provides that $\ell \propto z$. A frequently used form for ℓ is:

$$\frac{1}{\ell} = \frac{1}{\kappa z} + \frac{1}{\lambda}\,. \tag{10}$$

Here λ is a turbulent length scale, which should be valid for the turbulence far above the surface. We note that the latter has a rather empirical nature and consequently there is no agreement on the specification of λ in the literature.

Equation (8) is a diagnostic equation, which indicates that the eddy-diffusivity may variy with height, wind speed, stability, et cetera. In combination with the flux parameterization of Eq. (7), it follows that the flux at a certain height depends on the local gradient of the mean variable involved. Consequently the approach is referred to as a 'diagnostic local mixing approach'. Such an approach is mostly suitable for relatively homogeneous conditions with neutral and stable stratification, and is not so suitable for cases with convection (see non-local mixing parameterizations below).

More Advanced Mixing Parameterizations

A physically realistic alternative to the diagnostic approach is to relate the eddy-diffusivity of Eq. (7) to the actual turbulent kinetic energy of the flow, by using the prognostic turbulent kinetic energy equation and an appropriate choice for the turbulent length scale. It is important to realize that the kinetic energy of atmospheric motion per unit of mass E is given by the half of the sum of the velocities squared in the three directions (as in classic mechanics), e. g. $E = (U^2 + V^2 + W^2)/2$. Similar as with respect to Eq. (2), we can separate between the Mean Kinetic Energy $\overline{E}$ of the mean atmospheric motions and the Turbulent Kinetic Energy (TKE or e) of the smaller-scale fluctuating motions by turbulence. Thus e is given by $e = (\overline{u^2} + \overline{v^2} + \overline{w^2})/2$.

The prognostic equation for e reads in its basic form as:

$$\frac{\mathrm{D}e}{\mathrm{D}t} = -\overline{uw}\frac{\partial\overline{U}}{\partial z} - \overline{vw}\frac{\partial\overline{V}}{\partial z} + \frac{g}{\overline{\Theta_\mathrm{v}}}\overline{w\theta_\mathrm{v}} + D - \varepsilon\,. \tag{11}$$

Here $\mathrm{D}e/\mathrm{D}t$ is the total variation of e with time (the sum of local variations and those transported with the mean air motion). The two terms at the immediate right hand side of (11) represent the shear production of turbulence. These depend primarily on vertical variations of wind or, near the ground, on wind speed and surface roughness. The terms are almost always positive. The third term in Eq. (11) represents the rate of production or breakdown of turbulence by buoyancy effects (such as heat convection). It depends directly on density effects, which can be written in terms of the virtual potential temperature $\overline{\Theta_\mathrm{v}}$, and its turbulent flux $\overline{w\theta_\mathrm{v}}$. The term D in Eq. (11) represents divergence and pressure redistribution terms. These have a tendency to cancel near the surface. Finally, the term ε reflects the molecular dissipation of turbulence into heat and this term is always positive. In fact ε is typically proportional to e/τ, where τ is the characteristic time scale for the turbulent mixing process.

Using Eq. (11), turbulent kinetic energy can be calculated for given mean profiles when the corresponding fluxes are calculated using Eq. (7) for all fluxes involved. In this approach the diffusivities are typically calculated with equations of the form

$$K_c = \alpha_c \ell \sqrt{e}\,. \tag{12}$$

Here α_c is a constant depending on the variable of interest. The length scale is typically calculated with a similar type of diagnostic equation as (10) provides. This approach is known as the 'TKE-length scale approach' and it is an example of so-called 1.5 order closure. Sometimes a prognostic equation is used for the length scale as well, but such an approach is more popular in engineering applications then in the atmospheric sciences.

It can be shown that Eq. (8) is a solution of (11) and (12) in stationary conditions and when other simplifications are made such as the neglect of the influences by advection and turbulence divergence in the TKE equation. A more advanced turbulence scheme is known as 'second-order closure'. In such an approach, prognostic equations are developed for the fluxes and variances themselves. Such equations have a very similar structure as Eq. (12) for kinetic energy. Unfortunately, new unknowns are present in these equations. These must be related to the other variables in the model equations, always involving assumptions. Thus, second-order closure involves many more than the original equations and is therefore computationally more time consuming ('expensive') than first-order and 1.5 order closure.

One may expect that a model with 1.5 or second-order closure would produce more realistic results then a model with a first order closure. However, in practice this is often not the case, because of complex model interactions and the difficulty of representing all the relevant details with sufficient accuracy (see also below). That is the reason why diagnostic approaches remain popular. Nevertheless, second order equations are useful to gain insight in the governing physics, and after simplification useful extensions of the basic parameterizations may be achieved.

We continue our discussion with mixing parameterizations which have been proposed for boundary-layers with strong atmospheric convection. In such cases, the turbulent flux of a conserved quantity is typically not proportional to the local gradient alone as predicted by Eq. (7). In fact, in a large part of the ABL the mean gradients are small in conditions with dry convection, in particular for potential temperature (see Fig. 1). Then the fluxes depend mostly on the mixing characteristics of the large eddies across the ABL. Theories are available, which have modified K-theory to allow for the influence of convection, for example by including additional terms at the right hand side of Eq. (8) For details we refer to the literature (e. g., Holtslag and Moeng [15]).

In the next sections we apply the modeling concepts above and compare their results with field observations. In addition we present results from model intercomparison studies, and illustrate the role of boundary conditions.

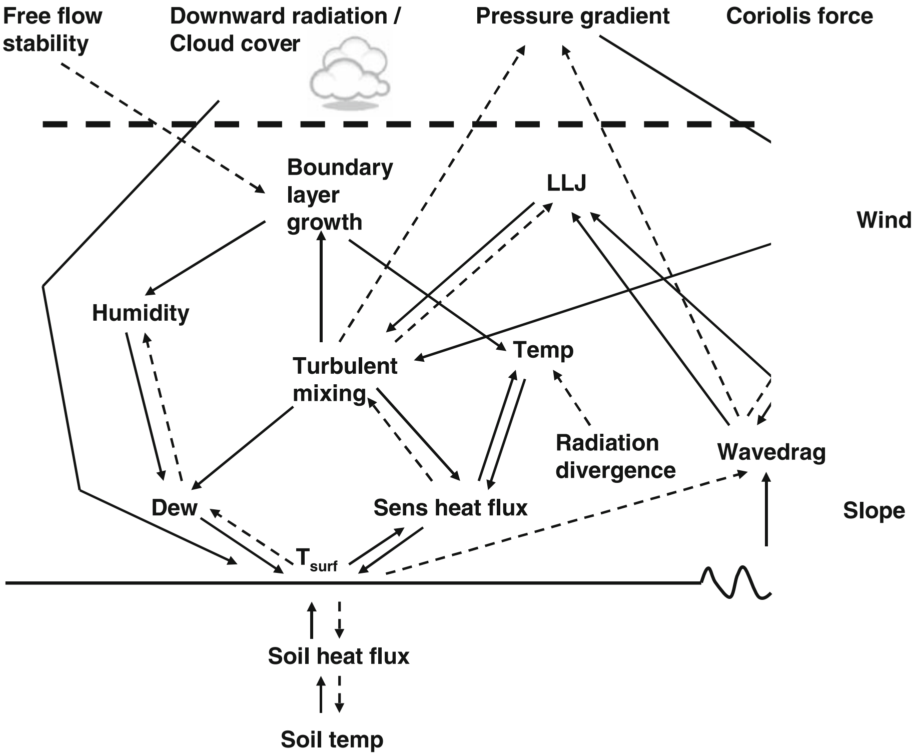
Intercomparison of Single Column Models for Stable Conditions

Atmospheric models for weather and climate need to make an overall representation of the smaller-scale boundary-layer and near surface processes. This appears to be more successful during daytime (e. g. Ek and Holtslag [10]; Holtslag and Ek [17]) then during nighttime stable conditions over land. The modeling of the stable boundary layer over land is rather complex because of the many different physical processes which are "at work" in stable conditions [21]. These small-scale processes are: clear air radiation divergence, drainage flow, generation of gravity waves and shear instabilities, fog and dew formation, the occurrence of a low-level jet and generation of discontinuous or intermittent turbulence [33]. In addition, the phenomenology of stable atmospheric boundary layers is quite divers, e. g. shallow and deep boundary layers with continuous turbulence through most of their depth, and on the other hand boundary layers with intermittent turbulence or even laminar flow.

The small-scale processes influence the vertical and horizontal exchange of quantities between the surface and the atmosphere as well as the mixing in the atmosphere on a variety of scales. In addition, it is known that turbulent mixing in stratified flow has an inherent non-linear character and may, as such, trigger positive feedbacks. These positive feedbacks, in turn, may cause unexpected transitions between totally different SBL regimes (e. g. van de Wiel et al. [34]).

Figure 3 depicts the interactions between relevant processes in the stable boundary layer. The non-linear behavior of the system is seen in e. g. the surface sensible heat-flux (H). A sudden change of the surface temperature can result in 2 different impacts on H. First, in weakly stable conditions (with strong wind and sufficient turbulence), a surface temperature decrease will provide a larger heat flux since H is proportional to the temperature difference between the surface and the atmosphere. The larger heat-flux from the atmosphere to the surface compsensates for the stronger cooling. In contrast for stronger stably stratified conditions, a surface temperature decrease will provide a stronger stratification and inhibits turbulent mixing, and consequently H will decrease. This allows for even stronger surface cooling (positive feedback). Note that a similar diagram for the daytime boundary layer can be found in Ek and Holtslag [10].

Having in mind the above mentioned complexity, one should not be surprised that atmospheric models encounter large forecast errors for stable conditions [20,24]. One strategy to improve model performance is to provide different models the same forecasting task, and analyze which model descriptions are in favor for which atmospheric stability. Recently such an intercomparison of boundary-layer schemes for stable conditions was made within the GEWEX Atmospheric Boundary Layer Study ('GABLS'). This GEWEX project aims to improve the understanding and the representation of the atmospheric boundary layer in regional and large-scale climate models [14]. A rather simple case was selected as a benchmark to review the state of the art and to compare the skills of single column (1D) models [7] and Large–Eddy Simulation models [3]. In this case a stable boundary layer is driven by an imposed, uniform geostrophic wind, with a specified constant surface-cooling rate over (homogeneous) ice. The case is initialized with $\theta = 265\,\mathrm{K}$ for $0 < z < 100$ and a lapse rate of 1 K/100 m aloft.



Single Column Modeling of Atmospheric Boundary Layers and the Complex Interactions with the Land Surface, Figure 3
Interaction diagram for the processes and variables in the stable atmospheric boundary layer over land (after [26])

It turns out that with the same initial conditions and model forcings, the models indicate a large range of results for the mean temperature and wind profiles. Figure 4 shows the mean profiles for several models after nine hours of constant surface cooling (sufficient to achieve a quasi-steady state). The variable results achieved are strongly related to the details of the boundary-layer mixing schemes [7]. An important finding is that the models in use at operational weather forecast and climate centers (as depicted at the left hand side of Fig. 4) typically allow for enhanced mixing resulting in too deep boundary layers, while the typical research models (at the right hand sides) show less mixing in more in agreement with the 'Large Eddy Simulation' results for this case [3].

Because of the enhanced mixing in weather and climate models, these models tend to show a too strong surface drag, too deep boundary layers, and an underestimation of the wind turning in the lower atmosphere [19]. At the other hand, by decreasing the mixing and surface drag, a direct impact on the atmospheric dynamics ('Ekman pumping') is noted (e. g. Beljaars and Viterbo [5]). Consequently, cyclones may become too active, corresponding in too high extremes for wind and precipitation, etc.

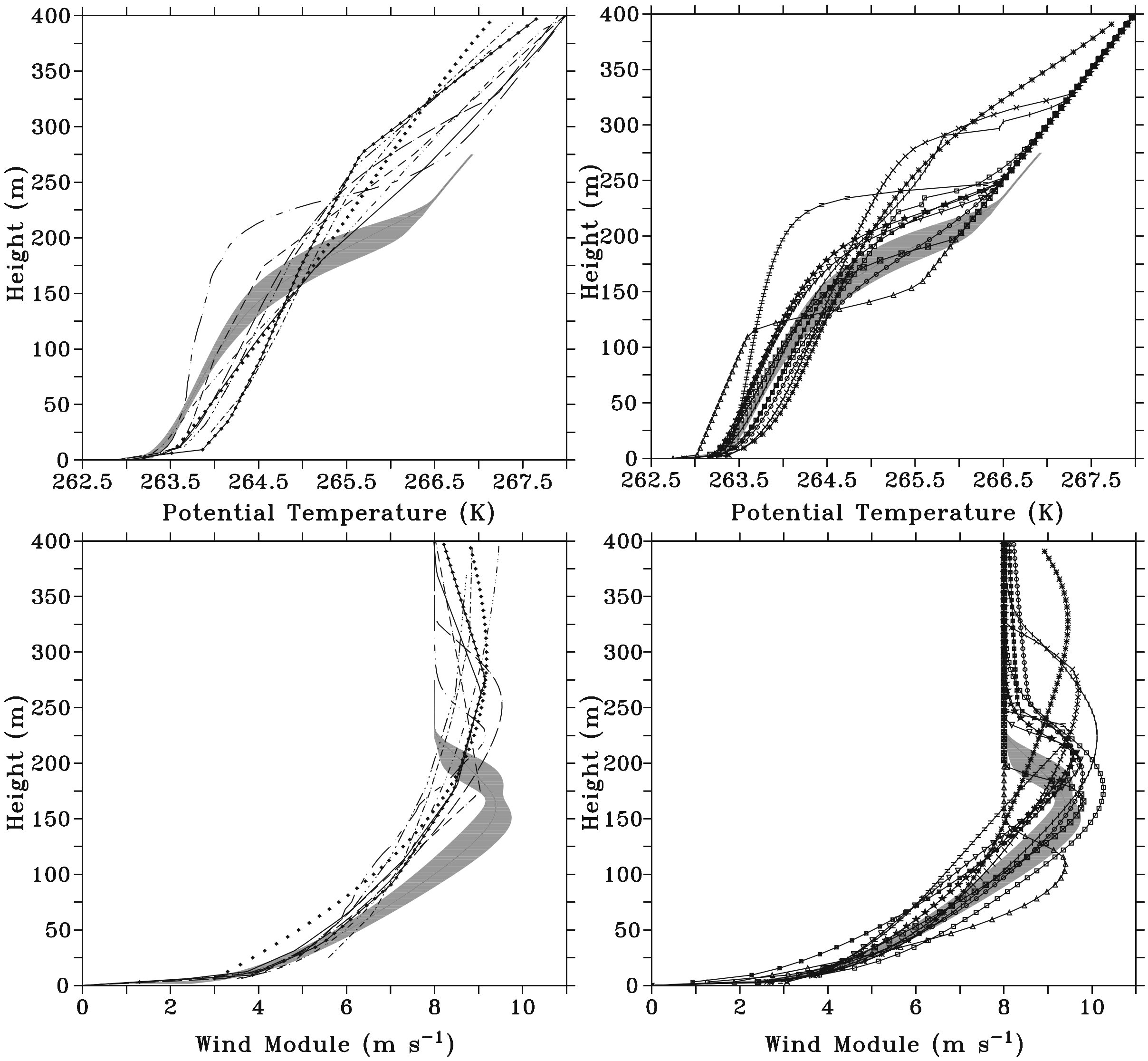
Modeling Boundary Layers over Land

To study the interactions of the ABL with the land surface we utilize the model by Duynkerke [9] with the extensions and modifications by Steeneveld et al. [28]. This model has been validated against tower observations at Cabauw, The Netherlands and later with CASES-99 field observations [1,28], and also participated in the GABLS model comparison in the previous section.

Instead of prescribing the surface temperature, and to enable interaction between the ABL and the land surface, the model is extended with a soil and a vegetation layer The soil temperature evolution is calculated by solving the diffusion equation (using a grid spacing of 1 cm) and the heat flux G_h from the soil to vegetation is calculated by:

$$G_h - (1 - f_{veg})K^{\downarrow} = r_g(T_{veg} - T_{s0})\,. \tag{13}$$

In Eq. (13) $K^{\downarrow}$ is the incoming shortwave (solar) radiation, T_{veg} represents the vegetation surface temperature, and T_{s0} the soil temperature just below the vegetation. We use a vegetation fraction $f_{veg} = 0.9$ and conductance $r_g = 5.9\,\mathrm{W m^{-2}} K^{-1}$, which are consistent with the obser-



Single Column Modeling of Atmospheric Boundary Layers and the Complex Interactions with the Land Surface, Figure 4
Modeled potential temperature and wind profiles by an ensemble of column models (after 9 h). *Grey areas* indicate the ensemble of Large Eddy Simulation results [3]. *Left panel* shows the results for first order closure models and the *right panel* for higher order closure models (after [7])

vations of CASES99 [28]. Initial soil and surface temperatures are also taken from the CASES99 observations.

Subsequently, the evolution of T_{veg} is computed by solving the surface energy budget for the vegetation layer:

$$C_{\text{v}} \frac{\partial T_{\text{veg}}}{\partial t} = Q^{*} - G_{\text{h}} - H - L_{\text{v}} E \, . \tag{14}$$

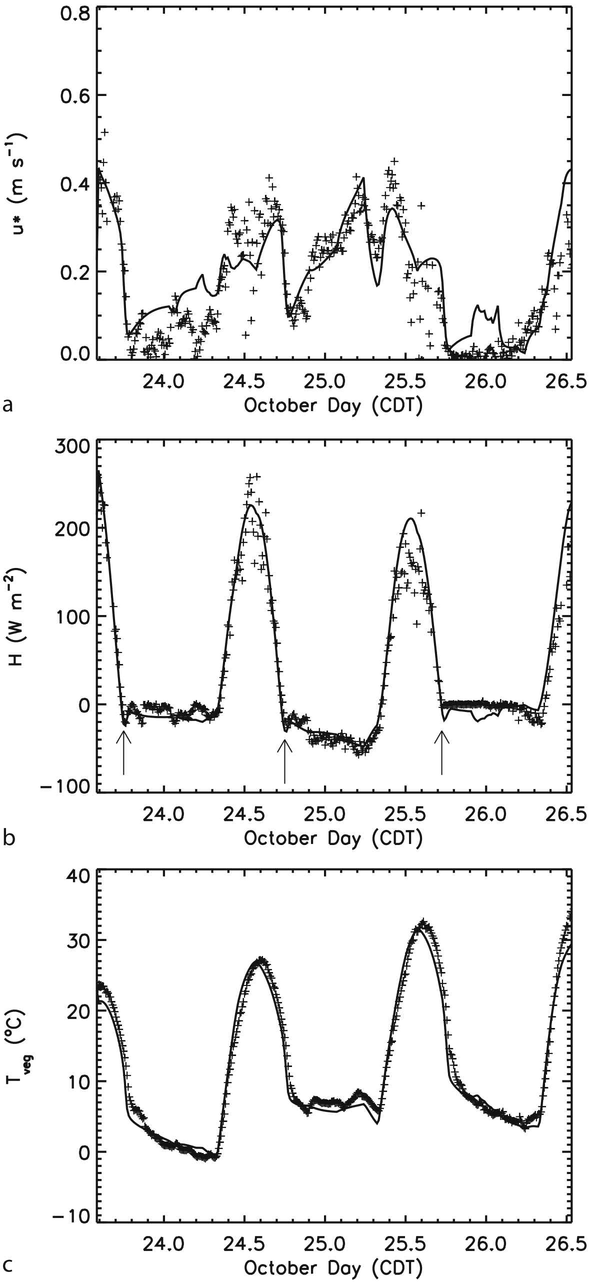
Here C_{v} is the heat capacity of the vegetation layer per unit of area ($C_{\text{v}} = 2000\,\text{Jm}^{-2}K^{-1}$, van de Wiel [34]), Q^{*} is the net radiation, H is the sensible heat flux and $L_{\text{v}}E$ the latent heat flux. The turbulent fluxes are calculated basically with the format of Eq. (6) above. Finally, Q^{*} is calculated by adopting the Garratt and Brost [12] radiation scheme. Note that Eqs. (13) and (14) provide a rather strong coupling of the atmosphere to the vegetated land surface for the current parameter setting which is found to be important [28].

Model forecasts have been compared with CASES-99 observations for contrasting diurnal cycles (i. e. for different wind speeds) for 23–26 Oct. 1999. Note that the model is only initialized once and that the total run comprises three full days. We distinguish between "radiative nights" with weak winds and small turbulent mixing, when radiative cooling dominates the SBL development. On the contrary, so-called "continuous turbulent" nights are characterized by strong winds. The final archetype is the so called "intermittently turbulent" night where turbulent episodes alternate with calm periods, when the radiative and turbulent forcings are of similar order of magnitude. Here we restrict ourselves to the results for surface fluxes, surface vegetation temperature, and vertical profiles of temperature and wind speed during nighttime. The diurnal cycle of the modeled net radiation Q^* (the balance of all incoming and outgoing short- and longwave radiative fluxes) agrees with the observations (not shown). Net radiation amounts typically $Q^* = 400\,\mathrm{Wm^{-2}}$ during daytime and $Q^* = -70\,\mathrm{Wm^{-2}}$ during nighttime.

The friction velocity (u_*) shows a clear diurnal cycle: u_* is large during the day and small at night, which is in general well captured by the model (Fig. 5a). Looking in more detail we find that during weak winds (1st and 3rd night) the model tends to overestimate u_*. The model lacks a clear turbulence collapse as observed during the first (intermittent) night. In the period 24 Oct., 700 CDT – 25 Oct., 1700 CDT the model performs well, while during the last (radiative) night u_* is slightly too high until midnight but follows the collapse at the end of the night. The overall bias amounts to $0.03\,\mathrm{ms^{-1}}$ for the last night. Sodar observations show much smaller wind speeds at 200 m AGL (which is above the SBL during this night) than the imposed **G**. This may suggest that **G** was overestimated, and this consequently may explain the bias in u_*. In general it is known that models correctly predict u_* for strong winds, but overestimate u_* for weak winds [25,32].

The sensible heat flux differs substantially between day ($\sim 250\,\mathrm{Wm^{-2}}$ here) and night (between 0 and $-60\,\mathrm{Wm^{-2}}$ depending on the wind speed). In the first (intermittent) night, the modeled $H = -14.1\,\mathrm{Wm^{-2}}$ on average, while –9.1 $\mathrm{Wm^{-2}}$ was observed. However, the model does not simulate the observed intermittent character of the surface fluxes (Fig. 5b). Some models with more resolution [23,36] were also able to reproduce intermittent turbulence. On the other hand, the models by Sharan and Gopalakrishnan [25] and Derbyshire [8] did not show any intermittency. This subject needs further investigation.

Just after the day-night transition to the intermittent night, the observed magnitude of H shows a clear maximum (see arrows in Fig. 5b), which is well reproduced by the model. This maximum is caused by a sudden reversal of the stratification near the surface due to longwave radiation emission during the day-night transition, and is maintained by residual turbulence of the convective boundary layer. This is an often observed *realistic feature* (e. g. during 11 of the 30 nights for CASES-99 and in FIFE observations shown by [6]). However, most model-



Single Column Modeling of Atmospheric Boundary Layers and the Complex Interactions with the Land Surface, Figure 5
Modeled and observed (+) friction velocity (a), surface sensible heat flux (b), and surface vegetation temperature (c) for three diurnal cycles in CASES-99 (after [28])

ing studies rarely show these minima. The reproduction of this detailed feature emphasizes the realism of the model outcome.

During the turbulent night (24–25 Oct.), the predicted H follows the observations. The specific minimum during the day-night transition (24 Oct., 1900 CDT) is present here as well. Radiative flux divergence dominates the last (radiative) night and the observed H is approximately zero. The model slightly overestimates the magnitude of H (–2.9 Wm^{-2}), mainly caused by an overestimation of u_*. This causes a weaker stratification and thus a larger magnitude of H. The second half of this night the model gives good results.

Reliable prediction of T_{veg} is a common problem for large-scale models. Some models show unphysical decoupling of the atmosphere from the surface resulting in so-called "runaway cooling" of T_{veg}. On the other hand, the pragmatic enhanced mixing approach which is commonly used for very stable conditions, leads to overestimation of T_{veg}. For both day- and nighttime T_{veg} *is simulated in very good agreement with the data, despite the fact that we cover a broad range of stability* (Fig. 5c).

We conclude that the present model generates surface fluxes which are in good agreement with observations, because of the detail in the description of the surface scheme, the soil heat flux and radiation physics (with high resolution). In general, the model is also able to estimate temperature and wind profiles (see results and discussion in Steeneveld et al. [28]). To examine the robustness of the results, we performed some sensitivity analysis on the initial conditions and model parameters. Disturbing the initial temperature (by 1 K), wind profiles (by 5%), soil temperature and vegetation temperature (both by 1 K) do not affect the results seriously. Also model re-initialization every 24 h (1400 CDT) with observed radiosonde information showed hardly any impact on the results (not shown).

Impact of Land Surface Conditions on Model Results

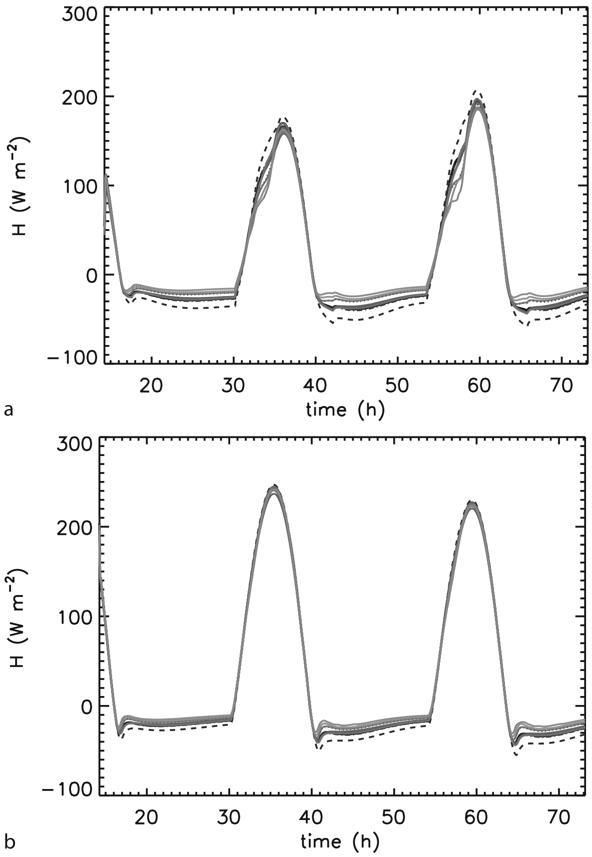
Inspired by the result in the previous section that a coupling with the land surface is necessary to obtain satisfactory model results, we now analyze the difference in variability of model results during model intercomparisons, as function of the chosen boundary condition. At first we use the first-order closure model and vary the parameters in the turbulence scheme for stable conditions in a reasonable range to mimic the apparent variability among boundary-layer models. As such first model runs are performed with a prescribed surface temperature as inspired by (but not identical to) the observations in CASES99 [22] and as described in the GABLS2 case description [31]. Subsequently, the model runs are repeated, but then using an interactive prognostic heat budget equation for the surface temperature (Eq. (14)).

To study the impacts of parameter values on the model results, reference runs are made for coupled and uncoupled cases with alternative permutations in some of the parameter settings for stable conditions. The parameter modifications are chosen such that they cover a realistic range in comparison with existing models of the stable boundary layer [7]. The local starting time in the model runs is 14.00 LT on October 22, 1999 (rather than 16.00 LT in the GABLS2 runs). The duration of all runs is 59 hours (so that the axis of all the figures indicates 14.00 until 73.00 h, covering a period of 2.5 diurnal cycles). In all model runs the roughness length for heat z_{oh} and momentum z_{om} (3 mm and 3 cm respectively), and the canopy resistance are constant, and the geostrophic wind is taken at a reference value of 9.5 ms^{-1} (as in Svensson and Holtslag [31]). The reference model set up has 50 logarithmically distributed layers and the first atmospheric model level is at 2 m.

The model results for all parameter permutations are presented for the sensible heat flux (Fig. 6), In the upper sub-frame of the figure (labeled a) the results achieved with the uncoupled model are given (using prescribed surface temperature). Overall the variety of results in the upper frame is comparable to the variety within the GABLS2 intercomparison study in stable conditions for the uncoupled models (see Svensson and Holtslag [31]). Thus we have a range of –15 to –50 Wm^{-2} for the sensible heat flux (at the end of the first night e. g. at the time of 30 h). The variability is a result of the range of parameters chosen above and the impact is apparently sufficient to mimic the different parameterizations for stable conditions in the models used within GABLS2.

Next we repeat all model runs and allow for surface feedback using Eqs. (13) and (14). The results for the sensible heat flux with the coupled model are given in the lower frame (Fig. 6b). Now we have a range of –10 to –25 Wm^{-2} (again the values apply for the end of the first night e. g. at the time of 30 h). Thus it appears that the variety of model results is *smaller* for the sensible heat flux in the coupled case. At the same time it appears that the variability appears to be somewhat larger for friction velocity and boundary-layer depth, which seems to be related to the larger variability in the near surface air temperature and wind speed.

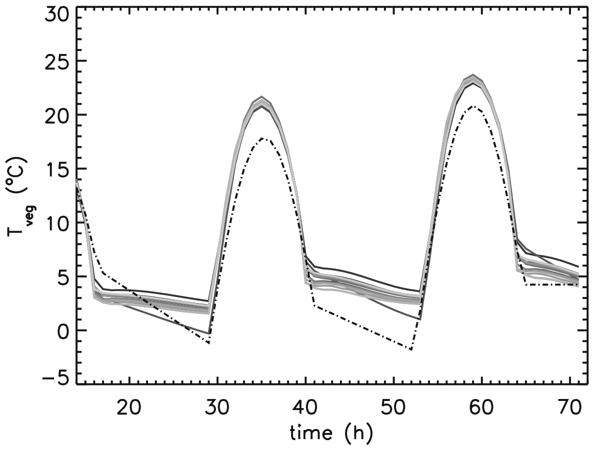
During daytime the sensible heat fluxes are rather similar for all model runs within one category (either coupled or uncoupled), but the maximum values differ. In addition, due to the coupling the sensible heat fluxes show



Single Column Modeling of Atmospheric Boundary Layers and the Complex Interactions with the Land Surface, Figure 6
Time series of an ensemble of model results for the sensible heat flux in a model intercomparison study with a prescribed surface temperatures, and b by solving the surface energy budget (after [18])

a more smooth behavior in the morning hours as compared with the uncoupled results. Thus, surface feedback is influencing the model results and is also able to compensate for some variation in the model parameter values. Note also that the variability in the friction velocities of the first night remains during the morning hours in the uncoupled runs, but not so much in the coupled case.

In Fig. 7 the surface temperatures are given as specified for the uncoupled case (the dashed line), and the temperatures as calculated in the various interactive model runs (various gray lines). It is seen that the latter ones are quite different from each other (in particular at night). It is also important to note that the surface temperature by the ensemble of coupled model runs is clearly different from the specified temperature in the uncoupled case. This impacts also on the absolute values and the range of air temperatures and the wind speeds.



Single Column Modeling of Atmospheric Boundary Layers and the Complex Interactions with the Land Surface, Figure 7
Time series of an ensemble of modeled surface temperature for coupled runs. *Dash-dotted line*: prescribed surface temperature in the uncoupled case (after [18])

Thus it is apparent that the treatment of the surface temperature impacts strongly on the outcome of the boundary-layer model results and their variety (see also Basu et al. [2]). By repeating the uncoupled model runs with a specified surface temperature as given by the ensemble mean value of the interactive runs, we achieve basically the same variety of model outputs for the potential temperature and wind as for the coupled cases. This confirms that in model evaluation studies the surface temperature should be taken consistent with the value of the geostrophic wind (although this may be model dependent).

Summary

In this paper a summary is given of the basic approaches for the modeling and parameterization of turbulence in the atmospheric boundary layer. The treated approaches are in current use in regional and global-scale models for the forecasting and study of weather, climate and air quality. Here we have shown the results of such approaches by using single column models in comparison with field data and fine-scale model results. We have also studied the impact of the surface temperature condition on the variability of results by an atmospheric boundary-layer model. From the coupled model results we achieve that surface feedback can compensate for some of the variety introduced by changing model parameters. Generally much work needs still to be done before we have a full understanding of the complexity of atmospheric turbulence and the interactions with the land surface. A better understanding of atmospheric turbulence hopefully also contributes to our capability in refining and unifying the turbulence parameterizations for modeling of the atmospheric boundary layer in response to the different type of surfaces which are found in reality.

Acknowledgments

This contribution is a compilation of earlier works by the authors, in particular the works by Holtslag [13], Steeneveld et al. [28] and Holtslag et al. [18].

Bibliography

1. Baas P, Steeneveld GJ, van de Wiel BJH, Holtslag AAM (2006) Exploring Self-correlation in flux-gradient relationships for stably stratified conditions. J Atmos Sci 63:3045–3054
2. Basu S, Holtslag AAM, van de Wiel BJH, Moene AF, Steeneveld GJ (2007) An inconvenienth 'truth' about using the sensible heatflux as a surface boundary condition in models under stably stratified regimes. Acta Geophys 56:88–99. doi:10.2478/s11600-007-0038-y
3. Beare R, MacVean M, Holtslag AAM, Cuxart J, Esau I, Golaz J-C, Jimenez M, Khairoutdinov M, Kosovic B, Lewellen D, Lund T, Lundquist J, McCabe A, Moene A, Noh Y, Raasch S, Sullivan P (2006) An intercomparison of Large–Eddy Simulations of the stable boundary layer. Bound-Layer Meteorol 118:247–272
4. Beljaars ACM, Holtslag AAM (1991) Flux parameterization over land surfaces for atmospheric models. J Appl Meteor 30:327–341
5. Beljaars ACM, Viterbo P (1998) Role of the boundary layer in a numerical weather prediction model. In: Holtslag AAM, Duynkerke PG (eds) Clear and Cloudy boundary layers. Royal Netherlands Academy of Arts and Sciences, Amsterdam, 372 pp
6. Chen F, Dudhia J (2001) Coupling an Advanced Land Surface-Hydrology Model with the Penn State-NCAR MM5 Modeling System. Part II: Preliminary Model Validation. Mon Wea Rev 129:587–604
7. Cuxart J, Holtslag AAM, Beare RJ, Bazile E, Beljaars A, Cheng A, Conangla L, Ek MB, Freedman F, Hamdi R, Kerstein A, Kitagawa H, Lenderink G, Lewellen D, Mailhot J, Mauritsen T, Perov V, Schayes G, Steeneveld GJ, Svensson G, Taylor P, Weng W, Wunsch S, Xu K-M (2006) Single-column model intercomparison for a stably stratified atmospheric boundary layer. Bound-Layer Meteorol 118: 273–303
8. Derbyshire SH (1999) Boundary layer decoupling over cold surfaces as a physical boundary instability. Bound-Layer Meteorol 90:297–325
9. Duynkerke PG (1991) Radiation fog: A comparison of model simulation with detailed observations. Mon Wea Rev 119:324–341
10. Ek MB, Holtslag AAM (2004) Influence of Soil Moisture on Boundary Layer Cloud Development. J Hydrometeor 5:86–99
11. Garratt JR (1992) The Atmospheric Boundary Layer. Cambridge University Press, New York, 316 pp
12. Garratt JR, Brost RA (1981) Radiative Cooling effects within and above the nocturnal boundary layer. J Atmos Sci 38:2730–2746
13. Holtslag AAM (2002) Atmospheric Boundary Layers: Modeling and Parameterization. In: Holton JR, Pyle J, Curry JA (eds) Encyclopedia of Atmospheric Sciences, vol 1. Academic Press, pp 253–261
14. Holtslag AAM (2006) GEWEX Atmospheric Boundary Layer Study (GABLS) on Stable Boundary Layers. Bound-Layer Meteorol 118:243–246
15. Holtslag AAM, Moeng C-H (1991) Eddy diffusivity and countergradient transport in the convective boundary layer. J Atmos Sci 48:1690–1698
16. Holtslag AAM, de Bruin HAR (1988) Applied Modeling of the Nighttime Surface Energy Balance over Land. J Clim Appl Meteor 27:689–704
17. Holtslag AAM, Ek MB (2005) Atmospheric Boundary Layer Climates and Interactions with the Land Surface. In: Encyclopedia of Hydrological Sciences. Wiley
18. Holtslag AAM, Steeneveld GJ, van de Wiel BJH (2007) Role of land-surface feedback on model performance for the stable boundary layer. Bound-Layer Meteorol 125:361–376
19. Lenderink G, Van den Hurk B, van Meijgaard E, van Ulden A, Cuijpers H (2003) Simulation of present day climate in RACMO2: First results and model developments. KNMI Technical report TR-252, 24 p
20. Mahrt L (1998) Stratified atmospheric boundary layers and breakdown of models. Theor Comp Fluid Phys 11:263–279
21. Mahrt L (1999) Stratified atmospheric boundary layers, Bound-Layer Meteorol 90:375–396
22. Poulos GS et al (2002) CASES-99: A comprehensive investigation of the stable nocturnal boundary layer. Bull Am Meteor Soc 83:555–581
23. ReVelle DO (1993) Chaos and "bursting" in the planetary boundary layer. J Appl Meteor 32:1169–1180
24. Salmond JA, McKendry IG (2005) A review of turbulence in the very stable boundary layer and its implications for air quality. Prog Phys Geogr 29:171–188
25. Sharan M, Gopalakrishnan SG (1997) Comparative Evaluation of Eddy Exchange coefficients for strong and weak wind stable boundary layer modelling. J Appl Meteor 36:545–559
26. Steeneveld GJ (2007) Understanding and prediction of stable boundary layers over land. Ph D thesis, Wageningen University, 199 pp
27. Steeneveld GJ, van de Wiel BJH, Holtslag AAM (2006) Modeling the arctic stable boundary layer and its coupling to the surface. Bound-Layer Meteorol 118:357–378
28. Steeneveld GJ, van de Wiel BJH, Holtslag AAM (2006) Modeling the Evolution of the Atmospheric Boundary Layer Coupled to the Land Surface for Three Contrasting Nights in CASES-99. J Atmos Sci 63:920–935
29. Steeneveld GJ, Mauritsen T, de Bruijn EIF, Vila-Guerau de Arellano J, Svensson G, Holtslag AAM (2008) Evaluation of limited area models for the representation of the diurnal cycle and contrasting nights in CASES99. J Appl Meteor Clim 47:869–887
30. Stull RB (1988) An introduction to Boundary-Layer Meteorology. Kluwer, Dordrecht, 666 pp. (reprinted 1999)
31. Svensson G, Holtslag AAM (2006) Single column modeling of the diurnal cycle based on CASES99 data -GABLS second intercomparison project. 17th Symposium on Boundary layers and Turbulence, 22–25 May, San Diego. American Meteorol Soc, Boston, Paper 8.1 (available at http://ams.confex.com/ams/pdfpapers)

32. Tjemkes SA, Duynkerke PG (1989) The nocturnal boundary layer: model calculations compared with observations. J Appl Meteor 28:161–175
33. van de Wiel BJH (2002) Intermittency and Oscillations in the Stable Boundary Layer over Land. Ph D thesis, Wageningen University, 129 pp
34. van de Wiel BJH, Moene AF, Hartogensis OK, de Bruin HAR, Holtslag AAM (2003) Intermittent turbulence and oscillations in the stable boundary layer over land, Part III: a classification for observations during CASES99. J Atmos Sci 60:2509–2522
35. van de Wiel BJH, Moene AF, Steeneveld GJ, Hartogensis OK, Holtslag AAM (2007) Predicting the Collapse of Turbulence in Stably Stratified Boundary Layers. Turbul Flow Combust 79:251–274
36. Welch RM, Ravichandran MG, Cox SK (1986) Prediction of Quasi-Periodic Oscillations in Radiation Fogs. Part I: Comparison of Simple Similarity Approaches. J Atmos Sci 43:633–651

Slug Flow: Modeling in a Conduit and Associated Elastic Radiation

Luca D'Auria, Marcello Martini
Osservatorio Vesuviano, Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Napoli, Naples, Italy

Article Outline

Glossary

Strombolian activity Is a kind of volcanic activity consisting of discrete, intermittent ejections of gas and magma fragments. The height reached by the ejecta seldom exceeds a few hundred meters above the volcanic crater. They consist mostly of molten magma fragments that partially quench as they fly. Strombolian explosions are intermediate between small Hawaiian eruptions and stronger Plinian eruptions. This kind of activity takes its name from the Stromboli volcano (Southern Italy) famous for its perpetual activity consisting of repeated moderate explosions (usually about 10 every hours).

Gas slug A bubble whose diameter is close to the diameter of the conduit where it is flowing. The shape and the motion of the slug is largely controlled by the conduit walls.

Modeling A simplified representation of a system aimed at reproducing some of its features. Mathematical models describe a system through a set of variables related by analytical relations. When these relations are too complex to be solved exactly, they can be solved using approximate numerical methods. Actually these numerical techniques often involve the massive use of computers. An alternative kind of modeling (analogue modeling) involves the use of versions of the system under study rescaled to fit spatial and temporal laboratory scales. Different materials are used to simulate original ones. For instance silicon oil is often used to simulate magma. The scaling relationships are rigorously stated in order to get physically meaningful results.

Computational fluid dynamics A set of mathematical, numerical and computational tools aimed at simulating complex fluid flows on computers. It developed simultaneously with computer science starting from the 1950s mostly with the aim of solving engineering problems. Today CFD spans a wider range of fields, from aeronautics to chemical engineering, to astrophysics to geophysics and much more. The most recent developments of CFD are related with the increase in computer performances and with the wider use of parallel computers.

Diffuse interface theory A molecular theory for describing the variation of the chemical composition across the interface on the basis of a rigorous thermodynamical approach. Starting from the definition of a free-energy function dependent on the chemical composition and its gradient, it is possible to compute all the chemical-physical properties of the interface. Beyond its physical meaning, this theory can be used also as a numerical tool for modeling of multiphase flows.

Very-long-period events (VLP) Seismic events recorded on active volcanoes and geothermal systems having a typical period of 10^2–10^0 s. Their observation and study began during the 1990s, with the spreading of seismic broadband sensors and have shown to be one of the most powerful tools for investigating the geometries of volcanic conduits and the dynamics of volcanic eruptions. Until now they have been observed in tens of volcanoes with different eruptive styles such as: Aso (Japan), Erebus (Antarctica), Kilauea (Hawaii), Popocatepetl (Mexico), Sakurajima (Japan), Stromboli (Italy).

Moment-tensor Tensor representation of the force systems acting on seismic sources. It can be applied to common earthquake sources (rupturing faults) as well as to volcanic sources (fluid filled conduits). In the former case its trace is null, which from a physical point of view, means that there is no netvolume change during common earthquakes. On the other hand in volcanic sources volumetric variations are very common and furthermore they are accompanied also by a single force component, related to the net acceleration of center of mass of the fluid filling the conduit.

Definition of the Subject

Among the eruptive styles, the Strombolian activity is one of the more easy to study because of its repetitive behavior. For this reason large amount of data can be comfortably collected. Strombolian volcanoes are like natural laboratories repeating the same experiment (individual explosions) many times each day.

The development of quantitative models of eruptive dynamics is driven by the comparison of experimental observations and synthetic data obtained through mathematical, numerical or analogue modeling.

Since Strombolian activity offers a profuse amount of interesting seismic signals, during the last decades there has been growing attention on seismological techniques aimed at retrieving the conduit geometry and the eruption dynamics from the seismological recordings. One of these techniques, the source function inversion, is able to retrieve a summary of the forces acting on the volcanic conduit during the VLP event generation [5]. The comparison of observed source functions with synthetic ones, obtained through numerical modeling, allow us to put constraints on the proposed models.

Quantitative models, able to fit seismological observations, are a powerful tool for interpreting seismic recordings and therefor the seismological monitoring of active volcanoes.

Introduction

In this paper we discuss the mechanism of generation of Very-Long-Period events related to Strombolian explosions. This eruptive style, occurring in many basaltic volcanoes worldwide, is characterized by the ascent and the bursting of large gas slugs. The mechanism of formation, ascent and explosion of bubbles and slugs and their relation to eruptive activity has been studied from a theoretical point of view and by means of analogue simulations. Here we introduce results from numerical simulations, focusing on the pressure variations induced on the conduit walls and responsible for the generation of seismic signals.

We will first illustrate the main features of the fluid dynamics related to Strombolian eruptive activity (Sect. "Slug Flow and Strombolian Activity") and an overview of the numerical modeling (Sect. "Numerical Modeling") Then we will show results obtained using simple conduit model (Sect. "Bubble Ascent", Sect. "Slug Ascent in a Vertical Pipe" and Sect. "Slug Ascent in a Pipe with a Flare") and we will compare the synthetic source functions with actual observations (Sect. "Seismological Constraints on Numerical Models").

Slug Flow and Strombolian Activity

A fundamental distinction can be made between eruptive regimes on the basis of the magma viscosity. In silicic systems, the magma viscosity ($>10^5$ Pa s) is too high to allow an independent motion of gas bubbles [18]. They can only grow by diffusion processes and expand under the effects of pressure variations until fragmentation occurs. In basaltic magmas the viscosity ($<10^3$ Pa s) allows independent motion of the gas bubbles leading to a different behavior with the possibility of bubble coalescence, splitting and turbulent fluid flow [11]. Theoretical laboratory and numerical studies have been published in order to understand the dynamics of Strombolian eruption in terms of gas/magma interaction [12,24].

The distinction between free bubble ascent and slug flow can be made using the ratio between the average bubble diameter d and the conduit diameter D [7]: the parameter λ (Table 1). Values of λ higher than 0.6 are related to slug flow. The range of behavior exhibited by slug flow depends on the physical properties of the fluids (density, viscosity, surface tension) and on the geometry of the conduit (diameter, shape, inclination). It is possible to define adimensional parameters for describing the particular flow regime (see Table 1). The Reynolds number is the ratio between inertial and viscous forces. Low values of Re are typical of laminar flows, while higher values are related to turbulent regimes. The Froude number Fr is the ratio between inertial and gravitational forces. The Eotvos number Eo is the ratio between buoyancy and surface tension effects. The values of Eo determines the bubble shape. High values of Eo are related to distorted bubble shapes, while lower values to sub-spherical shapes. The Morton number Mo has a similar meaning. The dimensionless inverse viscosity N_f assess the relative importance of viscous effects.

Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Table 1
Symbols used in the text

Symbol	Meaning
d	Average bubble diameter
D	Conduit diameter
L	Distance between the top of the bubble and the magma/air interface in the initial conditions.
U	Terminal slug velocity
λ	$\lambda = \frac{d}{D}$
g	Gravity
ρ	Liquid density
$\Delta\rho$	Difference between gas and liquid densities
μ	Dynamic viscosity
ν	Kinematic viscosity $\nu = \frac{\mu}{\rho}$
σ	Surface tension
γ	Isothermal expansion ratio
Re	Reynolds number $Re = \frac{Ud}{\nu}$
Fr	Froude number $Fr = \frac{U}{\sqrt{gD}}\sqrt{\frac{\rho}{\Delta\rho}}$
Mo	Morton number $Mo = \frac{g\mu^4\Delta\rho}{\rho^2\sigma^3}$
Eo	Eotvos number $Eo = \frac{\Delta\rho g D^2}{\sigma}$
N_f	Dimensionless inverse viscosity $N_f = \left(\frac{Eo^3}{Mo}\right)^{\frac{1}{4}}$

Viscous flows are characterized by $N_f < 2$ while inertial flows by $N_f > 200$ [9].

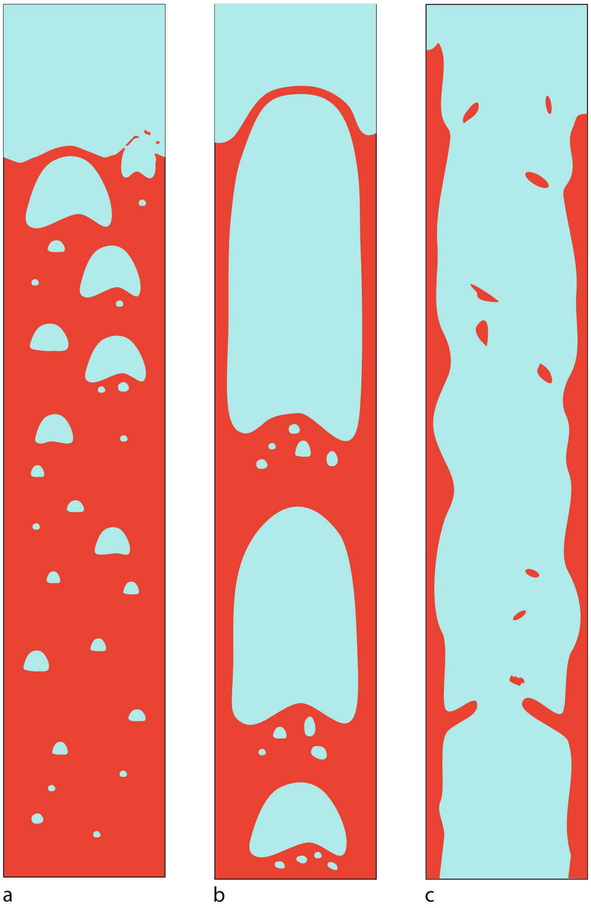
The ascent of gas in a basaltic system has been studied both from a theoretical and an experimental point of view. Observations of basaltic systems have shown that usually the flow conditions are transitional between viscous dominated and inertia dominated systems [17]. This puts constraints on the parameter range to explore both in numerical and analogue modeling. The range of adimensional numbers for basaltic systems, summarized from [17] and [9] is reported in Table 2.

For a basaltic system with given physical properties another variable plays a fundamental role in determining the physics of the flow: the gas/magma volumetric ratio. Flows are characterized by a low ratio consist of isolated bubbles rising with minor interaction between them and the conduit walls (Fig. 1a). In this regime bubbles rise assuming a shape that depends on their size and on the magma viscosity. Surface tension effect depends on the size of the bubbles and the ascent velocity depends on both the bubble size and the magma viscosity. In the range allowed for basaltic systems, larger bubbles have shapes ranging from dimpled ellipsoidal cap to spherical cap, while smaller bubbles have shapes ranging from spherical to ellipsoidal [7]. For higher gas/magma ratios bubbles begin to interact and to coalesce forming gas slugs occupying most of the conduit diameter (Fig. 1b). The ascent of gas slugs is strongly controlled by the conduit walls. During their passage, the magma is completely mixed and the motion of smaller bubbles is dominated by the fluid vorticity released (in a transitional regime) in their wake. For even higher gas/magma ratios (Fig. 1c) the slugs coalesce forming an almost continuous (turbulent) gas flow at the center of the conduit, while walls remains wet with a magma film. The high velocity of the flow produces instabilities on the surface of the magma film pulling out small fragments carried away from the gas.

Bubbly flows are related to effusive and to moderate Hawaiian explosive activity. Each single bubble explodes at the magma/air interface ejecting small magma fragments resulting from the shattering of the liquid film around the bubble. The bubble is usually overpressurized because of surface tension, viscous and inertial effects [23]. Slug flows are related instead to intermittent Strombolian activity.

Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Table 2
Range of adimensional numbers for flows related to basaltic systems (from [17] and [9])

Parameter	Range
Fr	$0.1 \div 0.345$
Re	$5 \div 31^3$
Eo	$51^5 \div 71^7$
Mo	$51^5 \div 101^{10}$
N_f	$16 \div 5000$



Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 1
Schematic representation of gas/magma flows in basaltic systems. a represents bubbly flow, b slug flow and c annular flow

Each slug exploding at the magma/air interface generates a jet of gas of short duration (from a few seconds to some tens of seconds) [24]. The main Strombolian explosion can be followed by minor bursts caused by the smaller bubbles trapped behind the slug wake. Annular flows are related to continuous lava fountains with the central gas jet carrying molten magma fragments [11].

The nature of the flow can change along the conduit because of the gas expansion due to the magmastatic pressure decrease. This expansion makes the gas/magma volumetric ratio increase along the conduit. At the base of the conduit, where the gas starts to exolve from the magma there is always a bubbly flow. As the pressure decreases the flow can change into slug flow. This can be made easier by an inclined conduit [4] and by constrictions [12] that force the bubbles to coalesce even at lower gas/magma ratios than those required in a straight vertical conduit. The total gas flow can change in the conduit because of variations in the deep feeding system or because of non-linear instabilities due to the complex shape of the conduit system [11,12].

The expansion of bubbles and slugs is a fundamental factor to take into account when studying Strombolian activity [17]. Assuming a simple ideal isothermal state equation for the gas we get the relative volumetric change for a bubble rising in a magma with density ρ:

$$\frac{V_0}{V} = \frac{P_{\text{atm}} + \rho g h}{P_{\text{atm}}} \,. \qquad (1)$$

It is easy to show that the expansion ratio in the upper few hundreds of meters is more than one order of magnitude.

The flow regime can change also horizontally, for instance in a subvertical dike the flow rate can be higher on one side, leading to a slug flow, and lower on the other side, leading to a bubbly flow. This explains the coexistence of effusive and Strombolian activity observed during some basaltic eruptions.

Numerical Modeling

The aim of this work is to investigate the pressure variations induced by a gas bubble rising in a magma-filled volcanic conduit. This phenomenon has been also investigated by means of analogue laboratory models [9,10, 12,17].

The major drawback of analogue modeling, in this context, is that it provides only a limited number of pressure time series: one for each sensor. Numerical modeling provides a different point of view giving the full set of scalar (density), vector (velocity) and tensor (pressure) quantities over the whole computational domain. This allows quantitative inferences on the flow regimes and on the elastodynamic wavefield generated (see Sect. "Seismological Constraints on Numerical Models").

The modeling of two-phase systems as gas magma is not a simple task in computational fluid dynamics (CFD) [6]. Taking into account surface tension effects at the liquid-gas interface can be done in two different ways. The first consists in explicitly tracking the time evolution of the gas-liquid interface [22]. In situations involving complex flows with extensive occurrences of bubble coalescence and splitting, these methods show a rapid increase in computational effort. Another category of methods model the gas-liquid systems using the diffuse-interface theory. These methods consider two scalar fields, defining the relative local concentrations of the two components and a modified advection-diffusion equation for

Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Table 3
Description and adimensional numbers for the six simulations presented in the text

Sim.	Description	Comp. domain (pts.)	λ	γ	*Re*	log(*Mo*)	*Eo*	*Fr*	N_f
1A	Bubble ascent (low viscosity)	60 × 600	0.5	5	1.93×10^3	0	9×10^4	0.37	5.19×10^3
1B	Bubble ascent (high viscosity)	60 × 600	0.5	5	1.90×10^2	4	9×10^4	0.36	5.19×10^2
2A	Slug ascent in vertical pipe (low viscosity)	30 × 600	0.8	5	1.82×10^3	0	9×10^4	0.35	5.19×10^3
2B	Slug ascent in vertical pipe (high viscosity)	30 × 600	0.8	5	1.54×10^2	4	9×10^4	0.30	5.19×10^2
3A	Slug ascent in vertical pipe with a flare (low viscosity)	60 × 600	0.8	5.3	1.83×10^3	0	9×10^4	0.35	5.19×10^3
3B	Slug ascent in vertical pipe with a flare (high viscosity)	60 × 600	0.8	5.3	1.68×10^2	4	9×10^4	0.32	5.19×10^2

modeling the physical-chemical interaction between them. This is done using a thermodynamically consistent definition of a free-energy function that takes into account the phase equilibria. This approach leads to smooth interfaces where the concentration of one phase gradually decreases while the other increases. The thickness of these interfaces depends on the numerical method and on the surface tension value [25]. From a numerical point of view this problem can be faced using Lattice Boltzmann Methods (LBM) [20] or by classic CFD [6]. Here we use the latter. Some details about the numerical method are reported in the Appendices.

In our models we have not considered the mechanical interaction between the fluid phases and the elastic conduit walls. The exchange of linear momentum between them is a significant factor in the generation of seismic waves [3] in seismo-volcanic sources. However in conduits having ratios between the linear dimension and thickness lower than 10^1–10^2 the effect of the motion of the conduit walls does not affect significantly the fluid flow. Furthermore we show in Sect. "Seismological Constraints on Numerical Models" that we will not compare the result of the simulation directly with seismograms, but with an equivalent system of forces acting on the conduit walls.

In the following we present results of some elementary two-dimensional conduit models focusing on the slug flow regime which occurs during Strombolian activity. The numerical simulations generate snapshots of the physical quantities (composition, density, pressure and velocity) along all the point of the discretized computational domain. The effective computational domains for each simulation are showed in Table 3. In all the simulations we start from a static fluid with a bubble in the lower part of the conduit and a gas/magma interface in the upper part. The boundary conditions keeps a constant pressure value at the bottom and the top of the model. The no-slip boundary conditions are implemented along the conduit walls.

The adimensional numbers for each simulation are reported in Table 3. Velocities for determining *Re* and *Fr* are computed when the bubble/slug is moving in a steady state after a transient due to the initial conditions. These velocity values U are also used for normalizing times:

$$t^* = t \frac{U}{L} , \tag{2}$$

where L is the distance between the top of the bubble and the magma/air interface. So (2) represents a normalization for the virtual time the bubble needs to reach the surface in an ideal steady motion. We can also normalize distances:

$$x^* = x \frac{1}{D} , \tag{3}$$

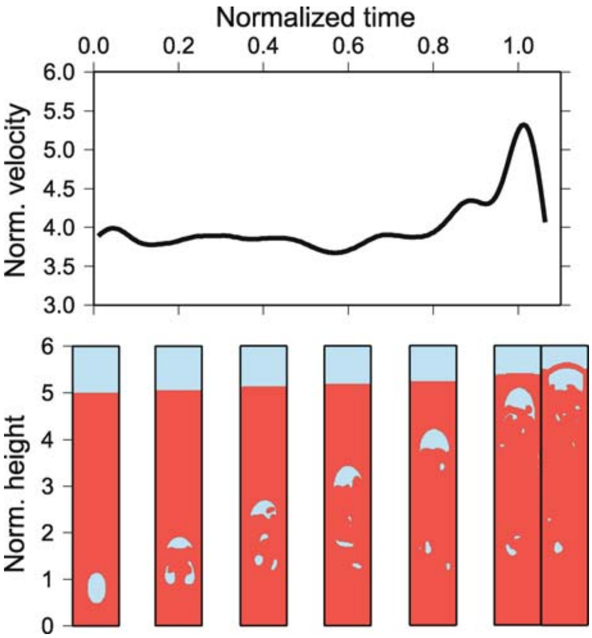
where D is the conduit diameter. We can define then normalized velocities as:

$$v^* = \frac{\partial x^*}{\partial t^*} . \tag{4}$$

In our simulations we have also defined an isothermal expansion ratio γ which is the ratio between the bubble volume at the initial conditions and its volume at (simulated) atmospheric pressure.

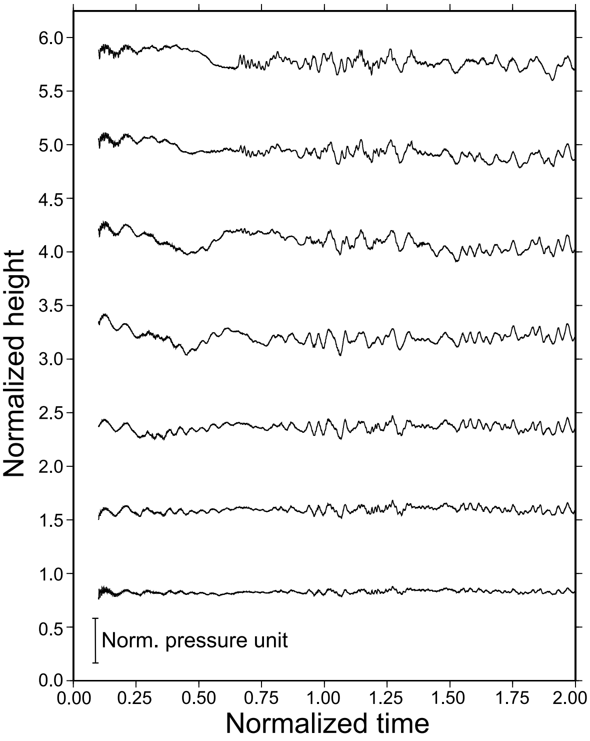
Bubble Ascent

We first consider the ascent of a single bubble in a vertical conduit. In the first simulation (1A in Table 3) the initial shape of the bubble is an ellipse (see Fig. 2). As the simulation evolves the bubble becomes unstable splitting into three smaller bubbles. The two smaller lateral bubbles are embedded in a symmetric vortex structure. As the flow



Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 2

Ascent and bursting of a gas bubble (simulation 1A in Table 3). Time, height and velocity are normalized according to (2), (3) and (4)



Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 3

Time series of pressure variations for simulation 1A. The represented pressure values are normalized by the external atmospheric pressure P_0. Each time series represents the recording of a virtual sensor located at the center of the conduit at the height of its first point (see *scale on the left*)

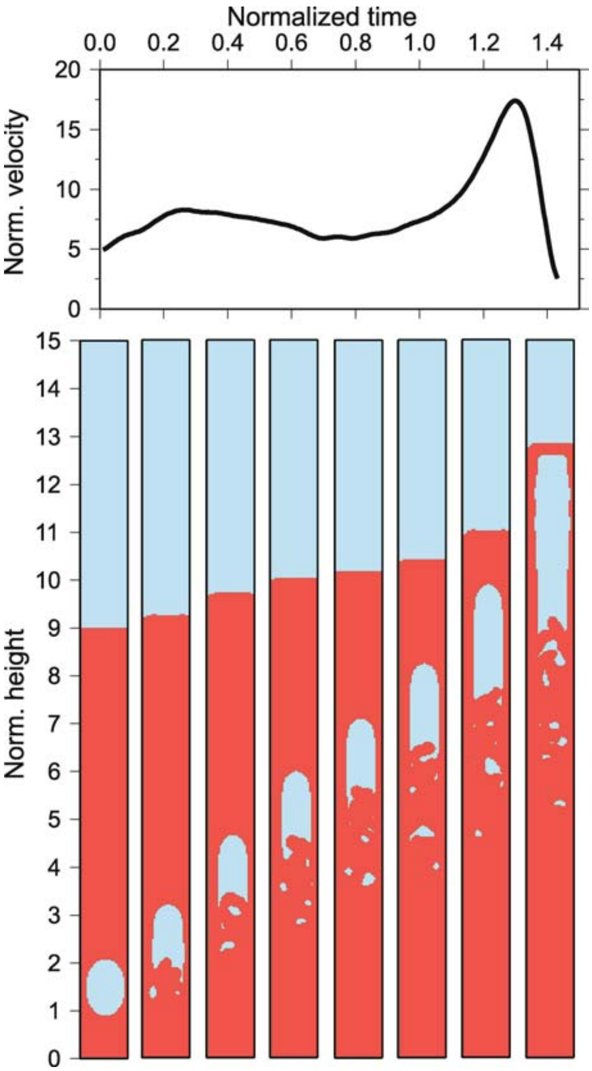
evolves the symmetry is broken and a turbulent wake develops behind the largest bubble. This one rises with an almost steady velocity (Fig. 2 top) with a spherical cap shape. In the final part of the simulation, the bubble suffers an expansion that made the bubble accelerate ($t > 0.8$) and causes the flow to become transitional toward the slug flow regime. Then the bubble reaches the surface and a curved liquid film develops above the bubble just before bursting. This phase is accompanied by a sudden deceleration of the bubble ascent ($t > 1.0$). The pressure variations in the conduit are modest, with a lack of low frequency oscillations. Also the pressure transients generated by the bubble burst are limited (Fig. 3).

The adimensional number for this simulation (Table 3) is in a range related to transitional flow regimes typical of basaltic systems [9,17]. An increase of an order of magnitude of the viscosity (1B in Table 3) makes the flow still in the transitional regime (Table 3). The high values of the Eotvos number indicates that the surface tension effect is not relevant in the flow dynamics [7] in both cases.

Slug Ascent in a Vertical Pipe

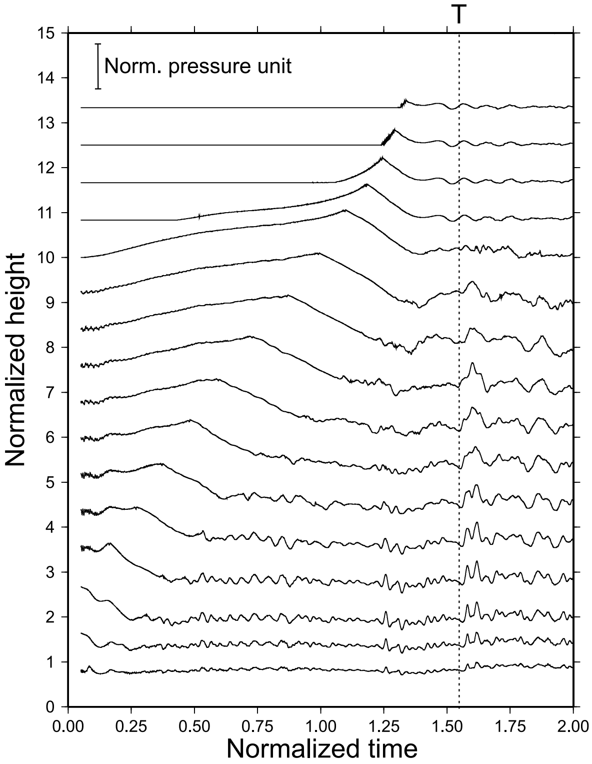
In these simulations (2A and 2B in Table 3) we model the ascent of a single slug in a vertical pipe. Again we start from an elliptical shape of the bubble. After an initial transient the typical slug shape develops (Fig. 4) and a turbulent wake appears behind the slug. As the slug rises it suffers a volume expansion. In the slug flow regime the liquid is pushed upward causing an overall increase of the hydrostatic pressure in the conduit. After the bursting of the slug the liquid film on the wall falls down to the original hydrostatic level. The slug expansion causes a significant acceleration of the liquid for $t > 1.0$ (Fig. 4). The velocity drops as the slug reaches the bursting point and the liquid film develops.

The major feature in the pressure pattern is the slow ramp-like increase followed by a drop (Fig. 5). The increase reflects the rise of the hydrostatic head above the slug while the gradual decrease is related to the passage of the slug. The bursting of the slug generates a moderate pressure transient in the conduit (marker “T” in Fig. 5) and damped resonant oscillations in the uppermost part of the pipe, filled with gas (above norm. height 11 in Fig. 5).



Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 4

Ascent and bursting of a single gas slug in a straight pipe (simulation 2A in Table 3)



Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 5

Time series of pressure variations for simulation 2A. *T* marks the most significant transient in the pressure patterns

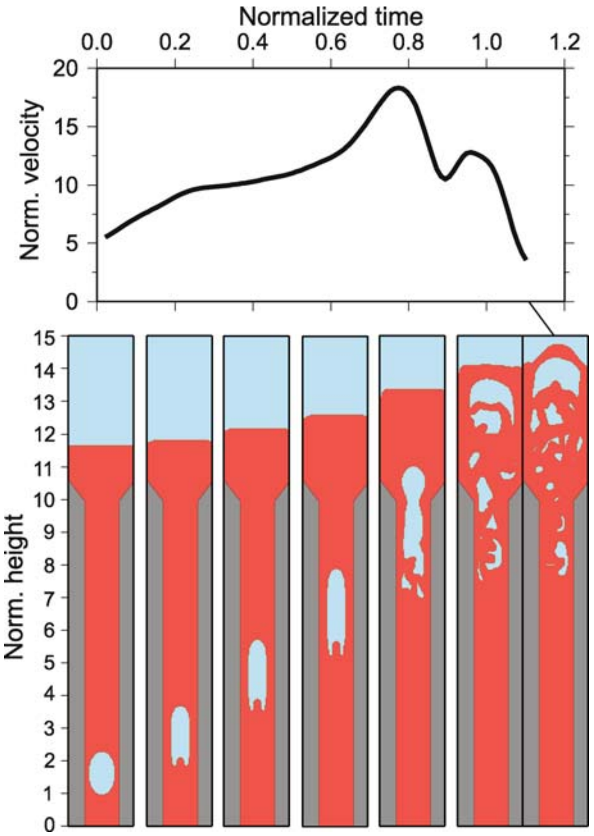
The simulated pressure pattern fits very well the observations of analogue modeling [9]. Both the pressure increase linked to the rise of the hydrostatic head and the pressure drop related to the passage of the slug are in good agreement. Together with these long period variations they observe also oscillatory pressure transients just before the slug approaches the surface and after the bursting. The simulations presented here lack this feature because of an intrinsic limit in the numerical method that does not model explicitly the interface (see Appendix).

The overall behavior of simulation 2B, where a tenfold increase of the liquid viscosity (Table 3) is very similar to 2A. The only remarkable difference is the lack of high frequency oscillations in the pressure (as the transient "T" in Fig. 5) due to the greater viscous damping effect.

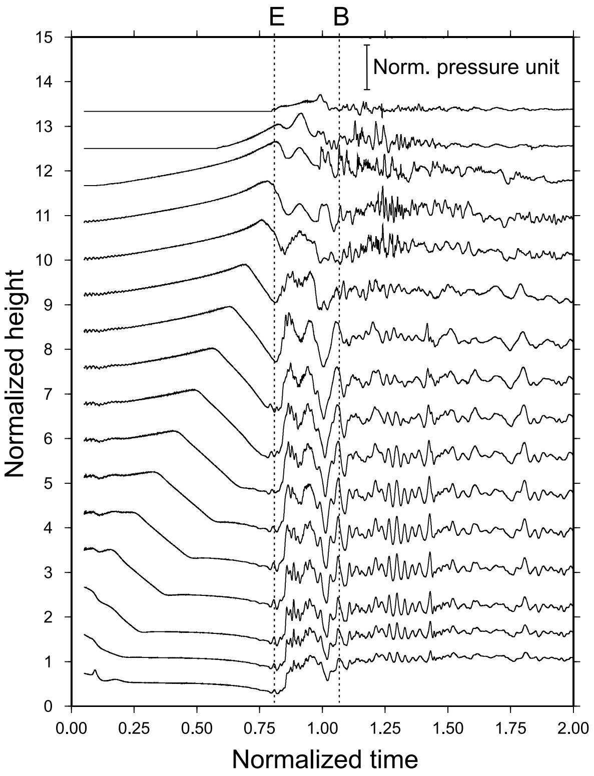
Slug Ascent in a Pipe with a Flare

In this set of simulations (3A and 3B in Table 3) we model again the ascent of a single slug in a pipe. However in this case the width of the pipe doubles after a norm. height of 10. In the first part of the simulation, the behavior is similar to the one illustrated in Sect. "Slug Ascent in a Vertical Pipe". As the slug nose enters in the flare (Fig. 6) is starts to expand rapidly making the fluid accelerate upward. The expansion of the slug in the flare is followed by its breakup because of the development of strong turbulence. The velocity of the top of the slug increases until the slug passes through the flare then it drops because of the change in the conduit diameter.

The pressure pattern shows a major difference, compared with simulation 2A (Fig. 7). As the slug passes through the flare there is a sudden pressure increase



Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 6
Ascent and bursting of a single gas slug in a pipe with a flare (simulation 3A in Table 3)



Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 7
Time series of pressure variations for simulation 3A. *E* marks the entering of the slug in the upper conduit while *B* marks the bursting of the main bubble at the magma surface

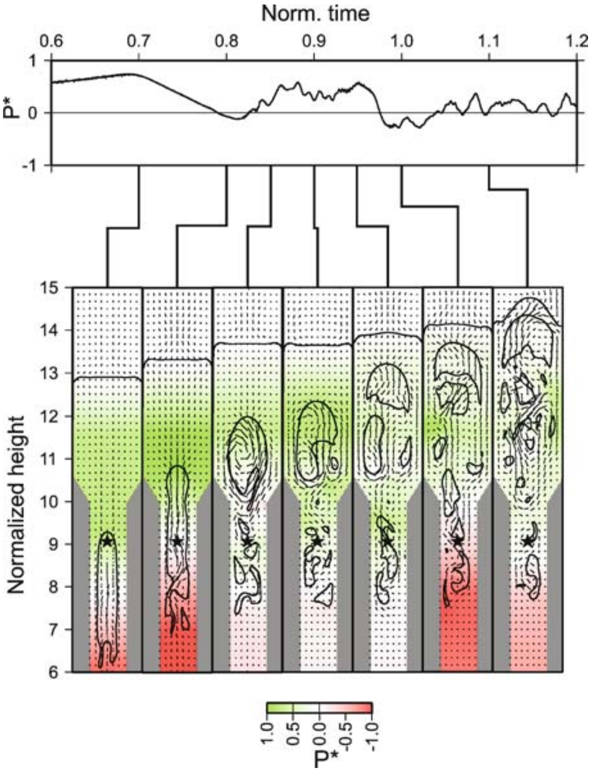
recorded along the whole conduit. The pressure then drops when the slug expansion has terminated. This strong pressure pulse is clearly related to the varying fluid flow regime as the slug enters the upper conduit and it deserves a closer analysis. In Fig. 8 we have represented the fluid dynamic regime during the slug expansion together with the related pressure variations. We observe that the pressure rises when most of the slug has passed through the flare. The strong acceleration induces turbulence both in the lower and the upper sections of the conduit. The upward acceleration first causes the disruption of the lower part of the slug leaving behind the main bubble a set of smaller ones, whose motion is driven by the fluid turbulence. When the slug has entered in the upper conduit it suffers another splitting due both to the induced vorticity and to the change in the boundary conditions. The behavior of the slug for $t^* > 0.8$ is characterized by a flow regime with a smaller value of λ and so it is similar to the initial part ($t^* < 0.2$) of simulation 1A.

Similar fluid dynamics behavior and pressure patterns have been observed in analogue simulations by [10]. In particular, the breakup of the slug and the generation of a positive pressure pulse as it passes through the flare are observed in analogue models within a wide range of fluid viscosities and conduit widenings.

As in the previous case the simulation with higher liquid viscosity (3B in Table 3) shows a similar behavior. Higher viscosities reduce the fluid vorticity leading to less fragmented slugs.

Seismological Constraints on Numerical Models

Seismological data analysis is a powerful tool for putting constraints on the geometry and the dynamics of volcanic systems. Non-stationary fluid flow in volcanic conduits generates pressure variations on the conduit walls and then seismic waves propagating toward the Earth's surface where they are recorded by seismometers. The frequency band of seismic signals recorded in volcanic ar-



Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 8
Detail of the conduit during the slug expansion phase. Each snapshot represents the normalized pressure variations (see the *color scale* on the bottom) and the local fluid velocity vectors. The *thick contours* are the bubble interfaces. The pressure values represented above are measured at the virtual sensor indicated by a *black star* in the snapshots

eas spans a wide range: from the Ultra-Long-Period (ULP) band ($>10^2$)s, the Very-Long-Period (VLP) band (10^2–10^0 s) and the Long-Period (LP) band (10^0–10^1 s) [5].

From a formal point of view this can be expressed considering the elastodynamic field generated by a general extended source, whose external surface is Σ:

$$u_n(\mathbf{x}, t) = \iint_\Sigma [f_q * G_{np} + m_{pq} * G_{np,q}] \mathrm{d}\Sigma \,, \quad (5)$$

where $u_n(\mathbf{x}, t)$ is the nth component of the ground displacement recorded at the position $\mathbf{x}$, f_q is the body-force distribution over Σ, m_{pq} is the moment density tensor and G_{np} are the Green's functions. In (5) we take into account all the details of the source dynamics. However the density of actual seismic networks is not sufficient for retrieving every minor feature of the seismic sources. Volcano monitoring networks usually have a limited extension (10^3–10^4 m) and the wavelengths associated with ULP and VLP signals are higher. Therefore the analysis of such signals the seismic source (i. e. the part of the volcanic conduit responsible of seismic wave generation) can be represented as a point [5]. Under the assumption of a point source we can express (5) as [5]:

$$u_n(\mathbf{x}, t) = F_q * G_{np} + M_{pq} * G_{np,q} \,, \quad (6)$$

where:

$$F_q = \iint_\Sigma f_q \mathrm{d}\Sigma \quad (7)$$

and:

$$M_{pq} = \iint_\Sigma m_{pq} \mathrm{d}\Sigma \,. \quad (8)$$

In common earthquake sources the single force component F_q is null. In volcanic sources, the acceleration of the center of mass of the fluid makes this component noteworthy.

The inversion of the recorder waveforms $\mathbf{u}(\mathbf{x}, t)$, after the numerical computation of the Green's functions $\mathbf{G}$ allows the retrieval of the single force and moment tensor components of (6) [14]. **F** and **M** are a synthesis of the force systems acting on volcanic conduits and then can be used for discriminating among numerical models on the basis of their fit with observations.

Numerical simulations provide the pressure tensor field **P** over the whole computational domain. The pressure tensor can be used for computing the forces acting on each point of the conduit walls multiplying it for the unit vector normal to the wall $\hat{\mathbf{n}}$:

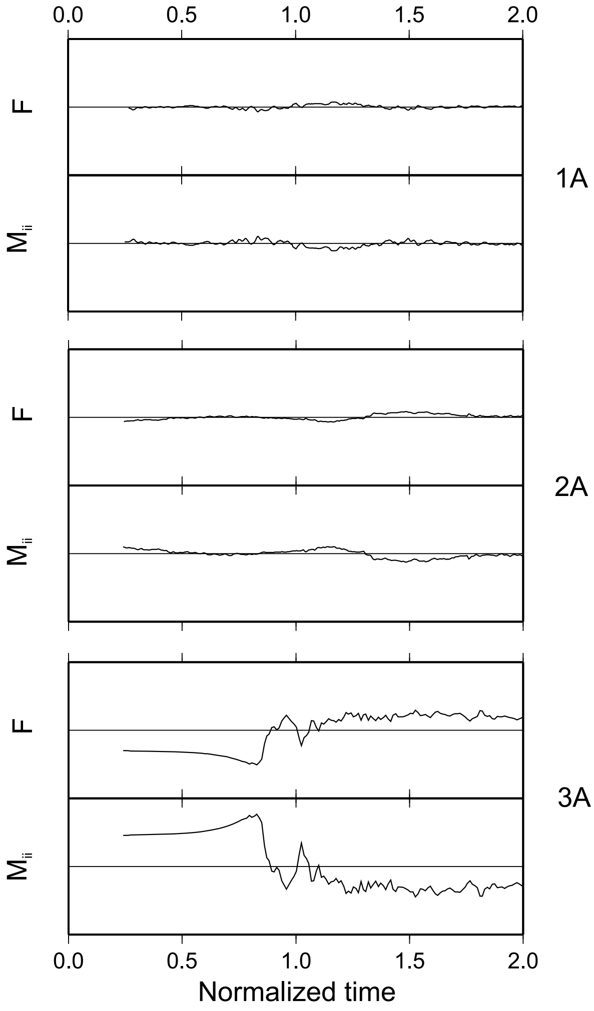
$$\mathbf{f} = \mathbf{P}\hat{\mathbf{n}} \,. \quad (9)$$

These values can be integrated numerically using an expression similar to (7) to get the single force component **F**. Then using the definition of moment they can be used for retrieving also the equivalent moment tensor:

$$M_{pq} = \iint_\Sigma \left(f_q - \frac{F_q}{\Sigma} \right) r_p \mathrm{d}\Sigma \,, \quad (10)$$

where **r** is the arm formed by the force respect to an arbitrary origin **O**.

In Fig. 9 we have represented the vertical force component and the moment tensor isotropic component (that is the trace $M_{xx} + M_{zz}$) for the three simulations with lower liquid viscosity. The represented values are normalized.



Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 9
Vertical force and isotropic component for simulations 1A, 2A and 3A of the moment tensor for simulations 1A, 2A and 3A (see Table 3)

In the first case (1A) we note the absence of significant signals. There are only minor oscillations related to fluid dynamics instability during the bubble ascent. The rise of a bubble smaller than the conduit dimension seems not to be able to generate VLP signals. Therefore a repetitive bursting of small bubbles at the magma/air interface can be responsible for the generation of continuous infrasonic and seismic tremor recorded in some basaltic volcanoes [16].

In the second case, again, the signal amplitudes are quite low. It is evident that there is an anticorrelation between the vertical force and the isotropic moment. The vertical acceleration of the slug ($1.0 < t^* < 1.3$ in Fig. 4) causes a downward reaction force as well as an increase in the conduit pressure. The bursting ($t^* > 1.3$) causes a downward acceleration of the liquid and a pressure decrease. A similar effect, although with a minor magnitude, is also evident in 1A (Fig. 9).

In the third case the signal amplitude increases dramatically, about five times higher than 2A. The slug entering in the upper, wider portion of the conduit is marked by a downward force and a positive moment. At the end of the expansion phase ($t^* > 0.8$) there is a sudden inversion of the trends in both quantities, followed again by a positive downward peak of the force and positive of the momentum ($1.0 < t^* < 1.1$). This peak occurs when two pieces of the original slug coalesce again (Fig. 8) leading to a sudden, but limited in time, upward acceleration. In real cases the dynamics can be even more complex with multiple slugs or slug fragments entering in the upper conduit and interacting with each other. In this case during the main expansion phase, the lower conduit is filled almost exclusively with liquid (Fig. 6). If we suppose the simultaneous presence of many vertically aligned slugs in the lower conduit, then the pressure variations induced by the expansion of the topmost one would influence the lower ones in a complex non-linear mechanism still to investigate.

The signals presented in Fig. 9 need to be scaled to an actual time scale to be compared with real seismic signals source functions. In basaltic volcanoes the dimensions of the upper conduit is of the order of 10^2 m while the ascent velocities of the slugs are of the order of 10^0–10^1 m/s. This gives a scaling factor of about 10^1–10^2 *s*. Since the simulated transients have a normalized duration of about 10^{-1} the actual simulated signals should have a characteristic period of 10^0–10^1 *s*. These values are within the range of VLP and LP signals [5].

The patterns observed in simulation 3A (Fig. 9) closely matches the results obtained by [4] for the source function of VLP events at the Stromboli volcano. Thus this result, together with analogue simulations [10], strongly supports the hypothesis that VLP events at Stromboli are generated by the passage of a gas slug through a conduit widening and its subsequent expansion and bursting. The long ramp-like signals in both force and moment for $t^* < 0.8$ in Fig. 9 has a characteristic period longer than the VLP signals recorded and analyzed by [4]. Recordings at Stromboli, using a seismic sensor with a wider frequency range [13] interestingly suggested in some seismic signals a similar ramp having a length of more than 60 *s*, compatible with our scaling. This long ramp is related to the slow and gradual increase in the magmastatic head due to the continuous expansion of the ascending slug.

Conclusions

The rapid expansion of the gas slugs in the uppermost part of the conduit plays a fundamental role, both in the eruptive dynamics and the seismic wave generation process. We have focused on the role of conduit geometry in the fluid dynamics of gas slug ascent and its implication in the generation of seismic signals. In one of the simulations (3A) we have shown that the system of forces acting on the conduit is able to generate seismic waves with a higher efficiency compared with other cases. This observation can be generalized to more complex geometry such as an alternating of widening and narrowing [10]. The passage of slugs can occur also in very complex conduits [17]. In this case it is possible that pressure transients are generated in different positions along the conduit.

Future Directions

A deeper understanding of the relationship between slug ascent dynamics and seismic signals generation would require more advanced modeling techniques in various three-dimensional geometries, testing how the effect of changes in slug volumes and magma properties can affect the generation of seismic signals.

These studies are an important step toward more advanced seismic monitoring techniques of active basaltic volcanoes aimed at assigning in real time a volcanological meaning to variations in observed LP and VLP seismic signals.

Appendix A – Fluid Dynamics of a Two-Phase System

Definition of a Two-Phase System

A fluid two-phase system can be described using two scalar fields n_a and n_b, representing the local molar densities of the two components a and b. The actual local density ρ is then:

$$\rho = m_\mathrm{a} n_\mathrm{a} + m_\mathrm{b} n_\mathrm{b} \,, \tag{11}$$

where m_a and m_b are the molecular weight of the two phases. As it will be shown in the following, it is convenient to describe the system using an alternative representation based on the variables:

$$n = n_\mathrm{a} + n_\mathrm{b} \,, \tag{12}$$

and

$$\phi = n_\mathrm{a} - n_\mathrm{b} \,. \tag{13}$$

ϕ defines the local composition of the fluid. It can span the range $[-n, +n]$ with the value $-n$ representing a pure b composition and $+n$ a pure a composition. The relation between n, ϕ and ρ is:

$$\rho = \tfrac{1}{2}\left[m_\mathrm{a}(n+\phi) + m_\mathrm{b}(n-\phi)\right] \,. \tag{14}$$

The values of m_a and m_b are set so that in reference conditions the value of n is always equal to 1. So for instance, if we consider pure water at ambient conditions $m_\mathrm{w} = 1000$ kg/mol.

State Equation

We assume that pure phases obey to a simple isothermal ideal gas state equation like:

$$P = \rho c^2 \,, \tag{15}$$

where P is the pressure and c is the sound speed. We assume that the pressure is zero in the reference state, so the state equation for the component a is:

$$P = m_\mathrm{a}(n-1)\,c_\mathrm{a}^2 \,. \tag{16}$$

A similar relation holds for b. Since along interfaces the composition varies continuously, we should define a state equation for a mixture that satisfies two requirements: the state equation for pure phases must match (16) and the isobaric contours in the (ϕ, n) plane must be straight lines. This second requirement follows from the consideration that diffusion processes between the components follows straight paths in this domain. In most of the actual Lattice Boltzmann literature, this question is not issued because the proposed implementations are almost isobaric $(n \simeq 1)$ [21,25,26]. A suitable choice for the state equation is then:

$$P(n,\phi) = \frac{1}{4}\left[\Lambda + \sqrt{16 m_\mathrm{a} m_\mathrm{b} c_\mathrm{a}^2 c_\mathrm{b}^2 (n-1) + \Lambda^2}\right] \tag{17}$$

with:

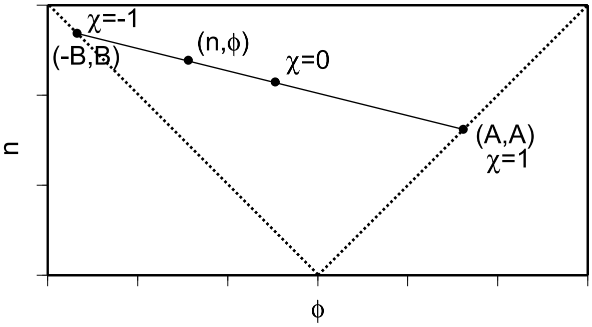
$$\Lambda = m_\mathrm{a} c_\mathrm{a}^2 (n+\phi-2) + m_\mathrm{b} c_\mathrm{b}^2 (n-\phi-2) \,. \tag{18}$$

In Fig. 11 an example contour plot of isobaric lines is shown. We can define another parameter χ as:

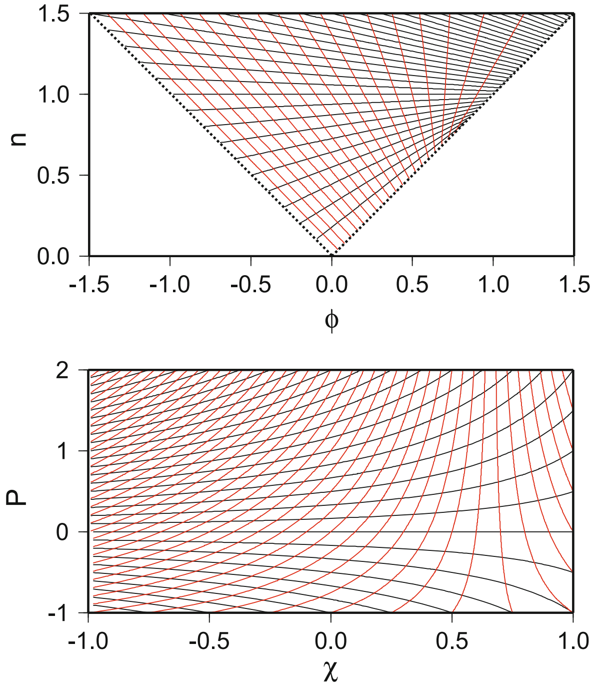
$$\chi(n,\phi) = -1 + 2\sqrt{\frac{(\phi+B)^2 + (n-B)^2}{(A+B)^2 + (A-B)^2}} \,, \tag{19}$$

where:

$$A = \frac{P(n,\phi)}{m_\mathrm{a} c_\mathrm{a}^2} + 1 \,, \tag{20}$$



Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 10
Definition of χ. See text for details



Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 11
In the upper panel, isolines of χ (*red lines*) and P (*black lines*) are represented in the (n, ϕ) domain. In the lower panel isolines of n (*black lines*) and ϕ (*red lines*) are represented in the (χ, P) domain. These plots represent a two-phase system where $m_a c_a^2 = 5$ and $m_b c_b^2 = 1$

and

$$B = \frac{P(n, \phi)}{m_b c_b^2} + 1 \,. \tag{21}$$

The actual meaning of χ is shown in Fig. 10. It is the rescaled ratio of the distance between the points (ϕ, n) and $(-B, B)$ and the distance between the points (A, A) and $(-B, B)$. So $\chi = -1$ along the line $\phi = -n$ and $\chi = 1$ along $\phi = n$. In practice χ is related to the local composition of the fluid. In Fig. 10 the relation between ϕ, n, P and χ is represented. This relation can be inverted and the system (χ, P) can be used as well as the (ϕ, n) coordinate system. In the following we show that the former is the best choice for describing chemical equilibria.

Chemical Equilibria and the Diffuse Interface Theory

The chemical equilibrium between the two phases can be described using a physically consistent thermodynamic approach [2,8,26]. We define a Ginzburg–Landau free energy functional for an heterogeneous mixture as [15]:

$$\mathcal{F} = \int \left[\psi(n, \phi) + \frac{\kappa_n}{2} (\nabla n)^2 + \frac{\kappa_\phi}{2} (\nabla \phi)^2 \right] \mathrm{d}V \,, \tag{22}$$

where κ_n and κ_ϕ are parameters related to the surface tension. The system reaches its final thermodynamical equilibrium making the free energy minimum.

We can rewrite (22) in the (χ, P) representation:

$$\mathcal{F} = \int \left[\psi(\chi, P) + \frac{\kappa_\chi}{2} (\nabla \chi)^2 \right] \mathrm{d}V \,. \tag{23}$$

Note that we have set $\kappa_P = 0$. This implies that surface tension effects are independent of the absolute pressure. In this paper we set:

$$\psi(\chi, P) = \alpha \left(\frac{\chi^4}{4} - \frac{\chi^2}{2} \right) + P \ln \frac{P}{P_0} \,, \tag{24}$$

where P_0 is an arbitrary reference pressure value. Using this definition of ψ we note that this function has two minima for $\chi = \pm 1$, corresponding to the two equilibrium compositions. Starting from an heterogeneous mixture, the system evolves by creating domains having a quite homogeneous composition separated by interfaces where the composition varies smoothly. The equilibrium is reached when the chemical potential:

$$\mu = \frac{\delta \mathcal{F}}{\delta \chi} \,, \tag{25}$$

is everywhere equal to zero [2,26]. So, in our case, using (23) and (24) we obtain the condition of chemical equilibrium for a plane interface orthogonal to the x-direction:

$$\mu = \frac{\partial \psi}{\partial \chi} - \kappa_\chi \frac{\partial^2 \chi}{\partial x^2} \,. \tag{26}$$

Assuming the boundary conditions $\psi(x) = 0$ and $\frac{\partial \chi}{\partial x} = 0$ for $x = \pm\infty$ and $\chi(0) = 0$, the previous ODE can be integrated giving the expression of the spatial variation of the composition through an equilibrium interface:

$$\chi(x) = \tanh\left(\frac{2x}{\xi}\right), \tag{27}$$

where the interface thickness ξ is the width where the 96% of variation occurs. Its value is:

$$\xi = 2\sqrt{-\frac{2\kappa_\chi}{\alpha}}. \tag{28}$$

On the basis of the definition of surface tension σ [25]:

$$\sigma = \int_{-\infty}^{+\infty} \mathcal{F}(x)\,\mathrm{d}x, \tag{29}$$

we can write:

$$\sigma = \frac{2}{3}\sqrt{-2\alpha\kappa_\chi}. \tag{30}$$

Using (28) and (30) we can also retrieve useful inverse relations:

$$\alpha = 3\frac{\sigma}{\xi}, \tag{31}$$

and

$$\kappa_\chi = \frac{3}{8}\sigma\xi. \tag{32}$$

We should emphasize that in real physical systems expression (27) describes an actual interface having a thickness whose order of magnitude is of molecular scale. In this work we use the diffuse interface theory as a numerical tool for modeling of two-phase systems. In other words, we use unphysical interfaces having a macroscopic thickness (usually $10^{-3} \div 10^{-1} m$).

Cahn–Hilliard and Mass Conservation Equations

The time evolution under non-equilibrium conditions can be expressed by two Cahn–Hilliard equations [2,26]:

$$\frac{\mathrm{D}n}{\mathrm{D}t} = \Gamma\nabla^2\mu_n, \tag{33}$$

$$\frac{\mathrm{D}\phi}{\mathrm{D}t} = \Gamma\nabla^2\mu_\phi. \tag{34}$$

These equations are similar to common advection-diffusion equations with Γ being a diffusion coefficient and the operator $\mathrm{D}/\mathrm{D}t$ the substantial derivative. The explicit expressions for the chemical potentials μ_n and μ_ϕ are:

$$\mu_n = \frac{\delta\mathcal{F}}{\delta n} = \frac{\delta\mathcal{F}}{\delta\chi}\frac{\partial\chi}{\partial n} = \mu_\chi\frac{\partial\chi}{\partial n}, \tag{35}$$

$$\mu_\phi = \frac{\delta\mathcal{F}}{\delta\phi} = \frac{\delta\mathcal{F}}{\delta\chi}\frac{\partial\chi}{\partial\phi} = \mu_\chi\frac{\partial\chi}{\partial\phi}. \tag{36}$$

with:

$$\mu_\chi = \frac{\delta\mathcal{F}}{\delta\chi} = \alpha\chi(\chi^2-1) - \kappa_\chi\nabla^2\chi. \tag{37}$$

Then the Cahn–Hilliard Eqs. (33) and (34) can be rewritten as:

$$\frac{\mathrm{D}}{\mathrm{D}t}\begin{pmatrix} n \\ \phi \end{pmatrix} = \Gamma\nabla^2\mu_\chi\begin{pmatrix} \frac{\partial\chi}{\partial n} \\ \frac{\partial\chi}{\partial\phi} \end{pmatrix} \tag{38}$$

On the basis of the definitions of n (12) and ϕ (13), we can state that the previous equation expresses the mass transfer of components a and b because of chemical disequilibria. Since we are dealing with a compressible flow, we should obviously account for this in the mass balances. Then, following basic fluid dynamics [1], we can rewrite (38) in explicit form as:

$$\frac{\partial n}{\partial t} = -\mathbf{v}\nabla n - n\nabla\cdot\mathbf{v} + \frac{\partial\chi}{\partial n}\Gamma\nabla^2\mu_\chi \tag{39}$$

$$\frac{\partial\phi}{\partial t} = -\mathbf{v}\nabla\phi - \phi\nabla\cdot\mathbf{v} + \frac{\partial\chi}{\partial\phi}\Gamma\nabla^2\mu_\chi. \tag{40}$$

Thermodynamic Pressure Tensor

It can be shown, from statistical mechanics that the pressure tensor can be obtained from (23) [8,15]:

$$P_{ij}^{th} = p_0\delta_{ij} + \kappa_\chi\frac{\partial\chi}{\partial x_i}\frac{\partial\chi}{\partial x_j}, \tag{41}$$

where the isotropic component is:

$$p_0 = P\frac{\delta\mathcal{F}}{\delta P} + \chi\frac{\delta\mathcal{F}}{\delta\chi} - \left(\psi(\chi,P) + \frac{\kappa_\chi}{2}(\nabla\chi)^2\right). \tag{42}$$

so:

$$p_0 = P + \alpha\left(\frac{3}{4}\chi^4 - \frac{1}{2}\chi^2\right) - \kappa\chi(\nabla^2\chi) - \frac{\kappa}{2}(\nabla\chi)^2 \tag{43}$$

The tensor $\mathbf{P}^{th}$ describes the stresses induced by spatial variation in density and composition and has to be added to the viscous stress tensor $\boldsymbol{\tau}$.

Conservation Equations

Let us now apply the results of the previous section for building a system of fluid dynamics equations suitable for a numerical implementation. Together with the mass conservation Eqs. (39) and (40) we need the conservation equation for the linear momentum [1]:

$$\rho \frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = \nabla \cdot \mathbf{P} + \rho \mathbf{g} \,, \tag{44}$$

where $\mathbf{P} = -\mathbf{P}^{th} + \boldsymbol{\tau}$ is the full pressure tensor of (41), $\mathbf{g}$ is the gravity and $\boldsymbol{\tau}$ is the viscous stress tensor that, for a Newtonian fluid is:

$$\tau_{ij} = \lambda e_{kk} \delta_{ij} + 2\mu e_{ij} \,, \tag{45}$$

with λ and μ the bulk and shear viscosities and e_{ij} is the strain rate tensor expressed by:

$$e_{ij} = \frac{1}{2} \left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right) \,. \tag{46}$$

Appendix B – Numerical Implementation

Transformed Equations

A common problem in the numerical solution of the conservation Eqs. (33) and (34) is the output of values having no physical meaning (for instance negative densities for one of the components) with bad effects on the numerical stability of the code. This is due to the fact that both variables are defined over a limited range of values that are $n \in [0, +\infty[$ and $\phi \in [-n, +n]$. These conditions are not explicit in the conservation equations, but this problem can be overcome by making simple changes of variable:

$$n = \mathrm{e}^{l} \,, \tag{47}$$

$$\phi = \mathrm{e}^{l} \tanh q \,. \tag{48}$$

Using the new variable l instead of n the conservation Eq. (39) become:

$$\frac{\partial l}{\partial t} = -\mathbf{v} \cdot \nabla l - \nabla \cdot \mathbf{v} + \frac{1}{\mathrm{e}^{l}} \left(\frac{\partial \chi}{\partial n} \right) \Gamma \nabla^2 \mu \,. \tag{49}$$

Substituting (48) in (40) we first obtain:

$$\frac{\partial q}{\partial t} = -\mathbf{v} \cdot \nabla q - \frac{1}{2} \sinh 2q \left(\frac{\partial l}{\partial t} + \mathbf{v} \cdot \nabla l + \nabla \cdot \mathbf{v} \right) + \frac{1}{\mathrm{e}^{l}} \cosh^2 q \left(\frac{\partial \chi}{\partial \phi} \right) \Gamma \nabla^2 \mu \,. \tag{50}$$

Using (49) and some algebraic manipulation the previous expression become:

$$\frac{\partial q}{\partial t} = -\mathbf{v} \cdot \nabla q - \frac{1}{\mathrm{e}^{l}} \cosh q \left[\frac{\partial \chi}{\partial n} \sinh q + \frac{\partial \chi}{\partial \phi} \cosh q \right] \cdot \Gamma \nabla^2 \mu \tag{51}$$

Boundary and Initial Conditions

Equations (49) and (51) have to be solved setting proper boundary and initial conditions.

The no-slip boundary condition, implemented along the conduit walls is:

$$\mathbf{v} = 0 \,. \tag{52}$$

Another boundary condition is set along the walls:

$$\nabla \phi \cdot \mathbf{n} = 0 \,. \tag{53}$$

This is the neutral wetting condition [25] and it is needed in order to avoid the wall to behave sticky respect to one of the phases.

At the conduit top and bottom conditions of constant pressure are implementer. On the top the pressure is kept to a reference atmospheric value, while at the bottom it is kept at the hydrostatic pressure value, computed on the initial conditions.

In the initial conditions we set $\mathbf{v} = 0$ and the computational domain is composed of domains having a homogeneous composition separated by smooth interfaces, close to the equilibrium solution of (27). At the top of the model a gas domain represents the atmosphere, while the remaining part of the conduit is filled with magma. A small elliptical gas bubble in hydrostatic equilibrium is placed in the lower part of the conduit. The details are specified for each set of simulations.

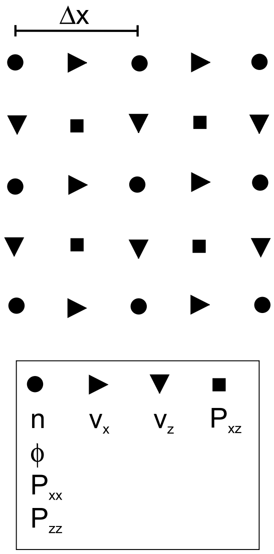
Finite Difference Implementation

Equations (49), (51) and (44) are discretized on a regular grid and the differential equations are transformed in finite-difference equations [19]. Scalar, vector and tensor quantities are discretized on staggered grids [6] (Fig. 12). Scalar quantities (l, q, μ and P_{ii}) are defined at integer grid steps (x_i, z_j). Velocities are staggered half grid step along x and z directions, that is v_x is defined at $(x_{i+1/2}, z_j)$ and v_z at $(x_i, z_{j+1/2})$. The deviatoric part of the pressure tensor P_{xz} is defined at grid points $(x_{i+1/2}, z_{j+1/2})$. This allows a second-order accuracy in the computation of spatial derivative.

We apply also a staggering in time. At the time step k we first solve (49) and (51). Then the current values of l and q are updated:

$$l_{i,j}^{(k)} = l_{i,j}^{(k-1)} + \frac{\mathrm{d}l}{\mathrm{d}t}^{(k)} \Delta t \,, \tag{54}$$

$$q_{i,j}^{(k)} = q_{i,j}^{(k-1)} + \frac{\mathrm{d}q}{\mathrm{d}t}^{(k)} \Delta t \,. \tag{55}$$



Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 12
Spatial scheme of the staggered grid used in the computation. *Circle* represents the isotropic grid (density, composition and isotropic pressure), *rightward triangles* are the grids for v_x while *downward triangles* are for v_z. *Squares* are the grids used for discretizing P_{xz}

Using these new values the full pressure tensor $\mathbf{P}^{(k)}$ is computed using the discretized version of (41), (42) and (43). At the time step $k + 1/2$ (44) is solved and the values of the local velocities are updated.

The computation of $\frac{dl}{dt}^{(k)}$ and $\frac{dq}{dt}^{(k)}$ requires the values of $l^{(k-1)}$, $q^{(k-1)}$ and $v^{(k-1/2)}$. On the other hand the computation of $v^{(k+1/2)}$ requires the values of $l^{(k)}$, $q^{(k)}$ and $v^{(k-1/2)}$.

The time step Δt is chosen in order to satisfy a stability condition. Since we are dealing with isothermal flows with a very low Mach number, the condition can be expressed by:

$$\Delta x < \Delta t \; c_{\max} \, , \qquad (56)$$

where $c_{\max}$ is the highest sound speed among the two components. In our simulations we have set:

$$\Delta x = 6 \; \Delta t \; c_{\max} \, , \qquad (57)$$

Bibliography

Primary Literature

1. Aris R (1962) Vectors, tensors and the basic equation of fluid mechanics. Dover Publications
2. Cahn JW, Hilliard JE (1958) Free energy of a nonuniform system. i. interfacial free energy. J Chem Phys 28(2):258–267
3. Chouet B (1986) Dynamics of a fluid-driven crack in three dimensions by the finite difference method. J Geophys Res 91:13967–13992
4. Chouet B, Dawson P, Ohminato T, Martini M, Saccorotti G, Giudicepietro F, De Luca G, Milana G, Scarpa R (2003) Source mechanisms of explosions at stromboli volcano, italy, determined from moment-tensor inversions of very-long-period data. J Geophys Res 108(B1)
5. Chouet BA (1996) New methods and future trends in seismological volcano monitoring. In: Scarpa R, Tilling RI (eds) Monitoring and mitigation of volcano hazards. Springer
6. Chung TJ (2002) Computational fluid dynamics. Cambridge University Press, Cambridge
7. Clift R, Grace JR, Weber ME (1978) Bubbles, drops and particles. Dover Publications
8. Evans R (1979) The nature of the liquid-vapour interface and other topics in the statistical mechanics of non-uniform, classical fluids. Adv Phys 28(2):143–200
9. James MR, Lane SJ, Chouet B, Gilbert JS (2004) Pressure changes associated with the ascent and bursting of gas slugs in liquid-filled vertical and inclined conduits. J Volc Geotherm Res 129:61–82
10. James MR, Lane SJ, Chouet BA (2006) Gas slug ascent through changes in conduit diameter: Laboratory insight into a volcano-seismic source process in low-viscosity magmas. J Geophys Res 111
11. Jaupart C (2000) Magma ascent at shallow levels. In: Sigurdsson H (ed) Encyclopedia of Volcanoes. Academic Press
12. Jaupart C, Vergniolle S (1988) Laboratory models of hawaiian and strombolian eruptions. Nature 331:58–60
13. Kirchdorfer M (1999) Analysis and quasistatic fe modeling of long period impulsive events associated with explosions at stromboli volcano (italy). Annali di Geofisica 42(3):379–390
14. Ohminato T, Chouet BA, Dawson P, Kedar S (1998) Waveform inversion of very long period impulsive signals associated with magmatic injection beneath kilauea volcano, hawaii. J Geophys Res 103(B10):23839–23862
15. Pooleay CM, Kuksenok O, Balazs AC (2005) Convection-driven pattern in phase-separating binary fluids. Phys Rev E 71:030501(R)
16. Ripepe M, Poggi P, Braun T, Gordeev E (1996) Infrasonic waves and volcanic tremor at stromboli. Geophys Res Lett 23(2):181–184
17. Seyfried R, Freundt A (2000) Experiments on conduit flow and eruption behaviour of basaltic volcanic eruptions. J Geophys Res B10(B10):23727–23740
18. Sparks RSJ (1978) The dynamics of bubble formation and growth in magmas: a review and analysis. J Volc Geotherm Res 3:137–186
19. Strikwerda JC (2004) Finite difference schemes and partial differential equations. SIAM, Philadelphia
20. Succi S (2001) The Lattice Boltzmann Equation for fluid dynamics and beyond. In: Numerical mathematics and scientific computation. Oxford University Press, Oxford
21. Swift MR, Orlandini E, Osborn WR, Yeomans JM (1996) Lattice Boltzmann simulations of liquid-gas and binary fluid system. Phys Rev E 54(5):5041–5052
22. Tomiyama A, Takagi S, Matsumoto Y (1999) Numerical simulation of bubble flows using interface tracking and bubble tracking method. Trans Model Simul 23

23. Vergniolle S, Brandeis G (1996) Strombolian explosions 1. a large bubble breaking at the surface of a lava column as a source of sound. J Geophys Res 101(B9):20433–20447
24. Vergniolle S, Mangan M (2000) Hawaiian and strombolian eruptions. In: Sigurdsson H (ed) Encyclopedia of Volcanoes. Academic Press
25. Xu A, Gonnella G, Lamura A (2003) Phase-separating binary fluids under oscillatory shear. Phys Rev E 67
26. Yue P, Feng JJ, Liu C, Shen J (2004) A diffuse-interface method for simulating two-phase flows of complex fluids. J Fluid Mech 515:293–317

Books and Reviews

Brennen CE (2005) Fundamentals of multiphase flow. Cambridge University Press, Cambridge

Scarpa R, Tilling RI (eds) (1996) Monitoring and mitigation of volcano hazards. Springer

Sigurdsson H, Bruce Houghton BF, McNutt SR, Rymer H, Stix J (eds) (2000) In: Encyclopedia of Volcanoes. Academic Press

Tannehill JC, Anderson DA, Pletcher RH (1997) Computational fluid mechanics and heat transfer, second edn. Taylor and Francis, London

Zobin V (2003) Introduction to volcanic seismology. Elsevier Science

Smooth Ergodic Theory

AMIE WILKINSON
Northwestern University, Evanston, USA

Article Outline

Glossary

Conservative, dissipative Conservative dynamical systems (on a compact phase space) are those that preserve a finite measure equivalent to volume. Hamiltonian dynamical systems are important examples of conservative systems. Systems that are not conservative are called dissipative. Finding physically meaningful invariant measures for dissipative maps is a central object of study in smooth ergodic theory.

Distortion estimate A key technique in smooth ergodic theory, a distortion estimate for a smooth map f gives a bound on the variation of the jacobian of f^n in a given region, for n arbitrarily large. The jacobian of a smooth map at a point x is the absolute value of the determinant of derivative at x, measured in a fixed Riemannian metric. The jacobian measures the distortion of volume under f in that metric.

Hopf argument A technique developed by Eberhard Hopf for proving that a conservative diffeomorphism or flow is ergodic. The argument relies on the Ergodic Theorem for invertible transformations, the density of continuous functions among integrable functions, and the existence of stable and unstable foliations for the system. The argument has been used, with various modifications, to establish ergodicity for hyperbolic, partially hyperbolic and nonuniformly hyperbolic systems.

Hyperbolic A compact invariant set $\Lambda \subset M$ for a diffeomorphism $f : M \to M$ is hyperbolic if, at every point in Λ, the tangent space splits into two subspaces, one that is uniformly contracted by the derivative of f, and another that is uniformly expanded. Expanding maps and Anosov diffeomorphisms are examples of globally hyperbolic maps. Hyperbolic diffeomorphisms and flows are the archetypical smooth systems displaying chaotic behavior, and their dynamical properties are well-understood. Nonuniform hyperbolicity and partial hyperbolicity are two generalizations of hyperbolicity that encompass a broader class of systems and display many of the chaotic features of hyperbolic systems.

Sinai–Ruelle–Bowen (SRB) measure The concept of SRB measure is a rigorous formulation of what it means for an invariant measure to be "physically meaningful". An SRB measure attracts a large set of orbits into its support, and its statistical features are reflected in the behavior of these attracted orbits.

Definition of the Subject

Smooth ergodic theory is the study of the statistical and geometric properties of measures invariant under a smooth transformation or flow. The study of smooth ergodic theory is as old as the study of abstract ergodic theory, having its origins in Bolzmann's Ergodic Hypothesis in the late 19th Century. As a response to Boltzmann's hypothesis, which was formulated in the context of Hamiltonian Mechanics, Birkhoff and von Neumann defined ergodicity in the 1930s and proved their foundational ergodic theorems. The study of ergodic properties of smooth systems saw an advance in the work of Hadamard and E. Hopf in the 1930s their study of geodesic flows for negatively curved surfaces. Beginning in the 1950s, Kolmogorov, Arnold and Moser developed a perturbative theory producing obstructions to ergodicity in Hamiltonian systems, known as Kolmogorov–Arnold–Moser (KAM) Theory. Beginning in the 1960s with the work of Anosov and Sinai on hyperbolic systems, the study of smooth ergodic theory has seen intense activity. This activity continues today, as the ergodic properties of systems displaying weak forms of hyperbolicity are further understood, and KAM theory is applied in increasingly broader contexts.

Introduction

This entry focuses on the basic arguments and principles in smooth ergodic theory, illustrating with simple and straightforward examples. The classic texts [1,2] are a good supplement.

The discussion here sidesteps the topic of ▶ Kolmogorov–Arnold–Moser (KAM) Theory, which has played an important role in the development of smooth ergodic theory. For reasons of space, detailed discussion of several active areas in smooth ergodic theory is omitted, including: higher mixing properties (Kolmogorov, Bernoulli, etc.), finer statistical properties (fast decay of correlations, Central Limit Theorem, large deviations), smooth thermodynamic formalism (transfer operators, pressure, dynamical zeta functions, etc.), the smooth ergodic theory of random dynamical systems, as well as any mention of infinite invariant measures. The text [3] covers many of these topics, and the texts [4,5,6] treat random smooth ergodic theory in depth. An excellent discussion of many of the recent developments in the field of smooth ergodic theory is [7].

This entry assumes knowledge of the basic concepts in ergodic theory and of basic differential topology. The texts [8] and [9] contain the necessary background.

The Volume Class

For simplicity, assume that M is a compact, boundaryless C^∞ Riemannian manifold, and that $f\colon M \to M$ is an orientation-preserving, C^1 map satisfying $m(D_x f) > 0$, for all $x \in M$, where

$$m(D_x f) = \inf_{v \in T_x M, \|v\|=1} \|D_x f(v)\| \,.$$

If f is a diffeomorphism, then this assumption is automatically satisfied, since in that case $m(D_x f) = \|D_{f(x)} f^{-1}\|^{-1} > 0$. For non-invertible maps, this assumption is essential in much of the following discussion. The Inverse Function Theorem implies that any map f satisfying these hypotheses is a covering map of positive degree $d \geq 1$.

These assumptions will avoid the issues of infinite measures and the behavior of f near critical points and singularities of the derivative. For most results discussed in this entry, this assumption is not too restrictive. The existence of critical points and other singularities is, however, a complication that cannot be avoided in many important applications. The ergodic-theoretic analysis of such examples can be considerably more involved, but contains many of the elements discussed in this entry. The discussion in Sect. "Beyond Uniform Hyperbolicity" indicates how some of these additional technicalities arise and can be overcome. For simplicity, the discussion here is confined almost exclusively to discrete time evolution. Many, though not all, of the the results mentioned here carry over to flows and semiflows using, for example, a cross-section construction (see Chap. 1 in [2]).

Every smooth map $f\colon M \to M$ satisfying these hypotheses preserves a natural measure *class*, the measure class of a finite, smooth Riemannian volume on M. Fix such a volume ν on M. Then there exists a continuous, positive *jacobian* function $x \mapsto \mathrm{jac}_x f$ on M, with the property that for every sufficiently small ball $B \subset M$, and every measurable set $A \subset B$ one has:

$$\nu(f(A)) = \int_B \mathrm{jac}_x f \, d\nu(x) \,.$$

The jacobian of f at x is none other than the absolute value of the determinant of the derivative $D_x f$ (measured in the given Riemannian metric). To see that the measure class of ν is preserved by f, observe that the Radon–Nikodym derivative $\frac{d f_* \nu}{d\nu}(x)$ at x is equal to $\sum_{y \in f^{-1}(x)} (\mathrm{jac}_y f)^{-1} > 0$. Hence $f_* \nu$ is equivalent to ν, and f preserves the measure class of ν.

In many contexts, the map f has a natural *invariant* measure in the measure class of volume. In this case, f is said to be *conservative*. One setting in which a natural invariant smooth measure appears is Hamiltonian dynamics. Any solution to Hamilton's equations preserves a smooth volume called the *Liouville measure*. Furthermore, along the invariant, constant energy hypermanifolds of a Hamiltonian flow, the Liouville measure decomposes smoothly into invariant measures, each of which is equivalent to the induced Riemannian volume. In this way, many systems of physical or geometric origin, such as billiards, geodesic flows, hard sphere gases, and evolution of the n-body problem give rise to smooth conservative dynamical systems. See ► Dynamics of Hamiltonian Systems.

Note that even though f preserves a smooth measure class, it might not preserve any measure in that measure class. Consider, for example, a diffeomorphism $f\colon S^1 \to S^1$ of the circle with exactly two fixed points, p and q, $f'(p) > 1 > f'(q) > 0$. Let μ be an f-invariant probability measure. Let I be a neighborhood of p. Then $\bigcap_{n=1}^{\infty} f^{-n}(I) = \{p\}$, but on the other hand, $\mu(f^{-n}(I)) = \mu(I) > 0$, for all n. This implies that $\mu(\{p\}) > 0$, and so μ does not lie in the measure class of volume. This is an example of a dissipative map. A map f is called *dissipative* if every f-invariant measure with full support has a singular part with respect to volume. As was just seen, if a diffeomorphism f has a periodic sink, then f is dissipative; more generally, if a diffeomorphism f has a periodic point p of period k such that $\mathrm{jac}_p f^k \neq 1$, then f is dissipative.

The Fundamental Questions

For a given smooth map $f\colon M \to M$, there are the following fundamental questions.

1. Is f conservative? That is, does there exist an invariant measure in the class of volume? If so, is it unique?
2. When f is conservative, what are its statistical properties? Is it ergodic, mixing, a K-system, Bernoulli, etc.? Does it obey a Central Limit Theorem, fast decay of correlations, large deviations estimates, etc.?
3. If f is dissipative, does there exist an invariant measure, not in the class of volume, but (in some sense) natural with respect to volume? What are the statistical properties of such a measure, if it exists?

There are several plausible ways to "answer" these questions. One might fix a given map f of interest and ask these questions for that specific f. What tends to happen in the analysis of a single map f is that either:

- the question can be answered using "soft" methods, and so the answer applies not only to f but to perturbations of f, or even to *generic* or *typical* f inside a class of maps; or
- the proof requires "hard" analysis or precise asymptotic information and cannot possibly be answered for a specific f, but can be answered for a large set of f_t in a typical (or given) parametrized family $\{f_t\}_{t\in(-1,1)}$ of smooth maps containing $f = f_0$.

Both types of results appear in the discussion that follows.

Lebesgue Measure and Local Properties of Volume

Locally, any measure in the measure class of volume is, after a smooth change of coordinates, equivalent to Lebesgue measure in $\mathbb{R}^n$. In fact, more is true: Moser's Theorem implies that locally any Riemannian volume is, after a smooth change of coordinates, *equal* to Lebesgue measure in $\mathbb{R}^n$. Hence to study many of the local properties of volume, it suffices to study the same properties for Lebesgue measure.

One of the basic properties of Lebesgue measure is that every set of positive Lebesgue measure can be approximated arbitrarily well in measure from the outside by an open set, and from the inside by a compact set. A consequence of this property, of fundamental importance in smooth ergodic theory, is the following statement.

Fundamental Principle #1: Two disjoint, positive Lebesgue measure sets cannot mix together uniformly at all scales.

As an illustration of this principle, consider the following elementary exercise in measure theory. First, some notation. If ν is a measure and A and B are ν-measurable sets with $\nu(B) > 0$, the *density of A in B* is defined by:

$$\nu(A\colon B) = \frac{\nu(A \cap B)}{\nu(B)} \, .$$

Proposition 1 *Let $\mathcal{P}_1, \mathcal{P}_2, \ldots$ be sequence of (mod 0) finite partitions of the circle S^1 into open intervals, with the properties: a) any element of $\mathcal{P}_n$ is a (mod 0) union of elements of $\mathcal{P}_{n+1}$, and b) the maximum diameter of elements of $\mathcal{P}_n$ tends to 0 as $n \to \infty$.*

Let A be any set of positive Lebesgue measure in S^1. Then there exists a sequence of intervals $I_1, I_2, \ldots$, with $I_n \in \mathcal{P}_n$ such that:

$$\lim_{n\to\infty} \lambda(A\colon I_n) = 1 \, .$$

Proof Assume that Lebesgue measure has been normalized so that $\lambda(S^1) = 1$. Fix a (mod 0) cover of $S^1 \setminus A$ by pairwise disjoint elements $\{J_i\}$ of the union $\bigcup_{n=1}^{\infty} \mathcal{P}_n$ with the properties:

$$\lambda(J_1) \geq \lambda(J_2) \geq \cdots \, , \text{ and}$$

$$\lambda\left(\bigcup_{i=1}^{\infty} J_i\right) = \sum_{i=1}^{\infty} \lambda(J_i) < 1 \, .$$

For $n \in \mathbb{N}$, let U_n be the union of all the intervals J_i that are contained in $\mathcal{P}_n$, and let $V_n = \bigcup_{i=1}^{n} U_n$. This defines an increasing sequence of natural numbers $i_1 = 1 < i_2 < i_3 < \cdots$ such that $U_n = \bigcup_{i=i_n}^{i_{n+1}-1} J_n$ and $V_n = \bigcup_{i=1}^{i_{n+1}-1} J_n$.

For each n, the interval I_n will be chosen to be an element $\mathcal{P}_n$, disjoint from V_n, in which the density of $\bigcup_{i=n+1}^{\infty} U_i$ is very small (approaching 0 as $n \to \infty$). Since $(S^1 \setminus A) \cap I_n$ is contained in $\bigcup_{i=n+1}^{\infty} U_i$, this choice of I_n will ensure that the density of A in I_n is large (approaching 1 as $n \to \infty$).

To make this choice of I_n, note first that the density of $\bigcup_{i=n+1}^{\infty} U_i$ inside of $S^1 \setminus V_n$ is:

$$\frac{\lambda(\bigcup_{i=n+1}^{\infty} U_i)}{\lambda(S^1 \setminus V_n)} = \frac{\sum_{i=i_{n+1}}^{\infty} \lambda(J_i)}{1 - \sum_{i=1}^{i_{n+1}-1} \lambda(J_i)} = a_n .$$

Note that, since $\sum_{i=1}^{\infty} \lambda(J_i)1$, one has $a_n \to 0$ as $n \to \infty$. Since the density of $\bigcup_{i=n+1}^{\infty} U_i$ inside of $S^1 \setminus V_n$ is at most a_n, there is an interval I_n in $\mathcal{P}_n$, disjoint from V_n, such that the density of $\bigcup_{i=n+1}^{\infty} U_i$ inside of I_n is at most a_n. Then

$$\lim_{n\to\infty} \lambda(A\colon I_n) \geq \lim_{n\to\infty} 1 - a_n = 1 .$$

□

In smooth ergodic theory, it is often useful to use a variation on Proposition 1 (generally, in higher dimensions) in which the partitions $\mathcal{P}_n$ are nested, dynamically-defined partitions. A simple application of this method can be used to prove that the doubling map on the circle is ergodic with respect to Lebesgue measure, which is done in Sect. "Lebesgue Measure and Local Properties of Volume".

Notice that this proposition does not claim that the intervals I_n are nested. If one imposes stronger conditions on the partitions $\mathcal{P}_n$, then one can draw stronger conclusions.

A very useful theorem in this respect is the Lebesgue Density Theorem. A point $x \in M$ is a *Lebesgue density point* of a measurable set $X \subseteq M$ if

$$\lim_{r\to 0} m(X\colon B_r(x)) = 1 ,$$

where $B_r(x)$ is the Riemannian ball of radius r centered at x. Notice that the notion of Lebesgue density point depends only on the smooth structure of M, because any two Riemannian metrics have the same Lebesgue density points. The Lebesgue Density Theorem states that if A is a measurable set and $\widehat{A}$ is the set of Lebesgue density points of A, then $m(A \,\Delta\, \widehat{A}) = 0$.

Ergodicity of the Basic Examples

This section contains proofs of the ergodicity of two basic examples of conservative smooth maps: irrational rotations on the circle and the doubling map on the circle. See ► Ergodic Theory: Basic Examples and Constructions for a more detailed description of these maps. These proofs serve as an elementary illustration of some of the fundamental techniques and principles in smooth ergodic theory.

Rotations on the circle. Denote by S^1 the circle $\mathbb{R}/\mathbb{Z}$, which is an additive group, and by λ normalized Lebesgue-Haar measure on S^1. Fix a real number $\alpha \in \mathbb{R}$. The rotation $R_\alpha\colon S^1 \to S^1$ is the translation defined by $R_\alpha(x) = x + \alpha$. Since translations preserve Lebesgue-Haar measure, the map R_α is conservative. Note that R_α is a diffeomorphism and an isometry with respect to the canonical flat metric (length) on S^1.

Proposition 2 *If $\alpha \notin \mathbb{Q}$, then the rotation $R_\alpha\colon S^1 \to S^1$ is ergodic with respect to Lebesgue measure.*

Proof Let A be an R_α-invariant set in S^1, and suppose that $0 < \lambda(A) < 1$. Denote by A^c the complement of A in S^1. Fix $\varepsilon > 0$. Proposition 1 implies that there exists an interval $I \subset S^1$ such that the density of A in I is large: $\lambda(A\colon I) > 1 - \varepsilon$. Similarly, one may choose an interval J such that $\lambda(A^c\colon J) > 1 - \varepsilon$. Without loss of generality, one may choose I and J to have the same length. Since α is irrational, R_α has a dense orbit, which meets the interval I. Since R_α is an isometry, this implies that there is an integer n such that $\lambda(R_\alpha^n(I) \,\Delta\, J) < \varepsilon\lambda(I)$. Since $\lambda(I) = \lambda(J)$, this readily implies that $|\lambda(A\colon R_\alpha^n(I)) - \lambda(A\colon J)| < \varepsilon$. Also, since A is invariant, and R_α is invertible and preserves measure, one has:

$$\lambda(A\colon R_\alpha^n(I)) = \lambda(R_\alpha^n(A)\colon R_\alpha^n(I)) = \lambda(A\colon I) > 1 - \varepsilon .$$

But for ε sufficiently small, this contradicts the facts that $\lambda(A\colon J) = 1 - \lambda(A^c\colon J) < \varepsilon$ and $|\lambda(A\colon R_\alpha^n(I)) - \lambda(A\colon J)| < \varepsilon$. □

Note that this is not a proof of the strongest possible statement about R_α (namely, minimality and unique ergodicity). The point here is to show how "soft" arguments are often sufficient to establish ergodicity; this proof uses no more about R_α than the fact that it is a transitive isometry. Hence the same argument shows:

Theorem 1 *Let $f\colon M \to M$ be a transitive isometry of a Riemannian manifold M. Then f is ergodic with respect to Riemannian volume.*

One can isolate from this proof a useful principle:

Fundamental Principle #2: Isometries preserve Lebesgue density at all scales, for arbitrarily many iterates.

This principle implies, for example, that a smooth action by a compact Lie group on M is ergodic along typical (nonsingular) orbits. This principle is also useful in studying area-preserving flows on surfaces and, in a refined form, unipotent flows on homogeneous spaces. In the case of surface flows, ergodicity questions can be reduced to a study of interval exchange transformations. See the entry ► Ergodic Theory: Basic Examples and Constructions for a detailed discussion of interval exchange transformations and flows on surfaces. ► Ergodic Theory, Introduction to contains detailed information on unipotent flows.

Doubling map on the circle. Let $T_2 \colon S^1 \to S^1$ be the doubling map defined by $T_2(x) = 2x$. Then T_2 is a degree-2 covering map and endomorphism of S^1 with constant jacobian $\mathrm{jac}_x T_2 \equiv 2$. Since $\frac{d(T_2)_*\lambda}{d\lambda} = \frac{1}{2} + \frac{1}{2} = 1$, T_2 preserves Lebesgue-Haar measure. The doubling map is the simplest example of a hyperbolic dynamical system, a topic treated in depth in the next section.

As with the previous example, the focus here is on the property of ergodicity. It is again possible to prove much stronger results about T_2, such as Bernoullicity, by other methods. Instead, here is a soft proof of ergodicity that will generalize readily to other contexts.

Proposition 3 *The doubling map $T_2 \colon S^1 \to S^1$ is ergodic with respect to Lebesgue measure.*

Proof Let A be a T_2-invariant set in S^1 with $\lambda(A) > 0$. Let $p \in S^1$ be the fixed point of T_2, so that $T_2(p) = p$. For each $n \in \mathbb{N}$, the preimages of p under T_2^{-n} define a (mod 0) partition $\mathcal{P}_n$ into 2^n open intervals of length 2^{-n}; the elements of $\mathcal{P}_n$ are the connected components of $S^1 \setminus T_2^{-n}(\{p\})$. Note that the sequence of partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots$ is nested, in the sense of Proposition 1. Restricted to any interval $J \in \mathcal{P}_n$, the map ${T_2}^n$ is a diffeomorphism onto $S^1 \setminus \{p\}$ with constant jacobian $\mathrm{jac}_x(T_2^n) = (T_2^n)'(x) = 2^n$.

Since A is invariant, it follows that $T_2^{-n}(A) = A$. Fix $\varepsilon > 0$. Proposition 1 implies that there exists an $n \in \mathbb{N}$ and an interval $J \in \mathcal{P}_n$ such that $\lambda(A \colon J) > 1 - \varepsilon$. Note that $T_2^n(A \cap J) \subset A$. But then

$$\begin{aligned}\lambda(A) &\geq \lambda(T_2^n(A \cap J)) \\ &= \int_{A \cap J} \mathrm{jac}_x(T_2^n)\, \mathrm{d}\lambda(x) \\ &= 2^n \lambda(A \cap J) \\ &= 2^n \lambda(A \colon J)\lambda(J) \\ &> 2^n(1-\varepsilon)\lambda(J) = 1 - \varepsilon\,.\end{aligned}$$

Since ε was arbitrary, one obtains that $\lambda(A) = 1$. □

In this proof, the facts that the intervals in $\mathcal{P}_n$ have constant length 2^{-n} and that the jacobian of ${T_2}^n$ restricted to such an interval is constant and equal to 2^n are not essential. The key fact really used in this proof is the assertion that the ratio:

$$\frac{\lambda(T_2^n(A \cap J) \colon T_2^n(J))}{\lambda(A \colon J)}$$

is bounded, *independently of n*. In this case, the ratio is 1 for all n because T_2 has constant jacobian.

It is tempting to try to extend this proof to other expanding maps on the circle, for example, a C^1, λ-preserving map $f \colon S^1 \to S^1$ with $d_{C^1}(f, T_2)$ small. Many of the aspects of this proof carry through *mutatis mutandis* for such an f, save for one. A C^1-small perturbation of T_2 will in general no longer have constant jacobian, and the *variation* of the jacobian of f^n on a small interval can be (and often is) unbounded. The reason for this unboundedness is a lack of control of the modulus of continuity of f'. Hence this argument can fail for C^1 perturbations of T_2. On the other hand, the argument still works for C^2 perturbations of T_2, even when the jacobian is not constant.

The principle behind this fact can be loosely summarized:

Fundamental Principle #3: On controlled scales, iterates of C^2 expanding maps distort Lebesgue density in a controlled way.

This principle requires further explanation and justification, which will come in the following section. The C^2 hypothesis in this principle accounts for the fact that almost all results in smooth ergodic theory assume a C^2 hypothesis (or something slightly weaker).

Hyperbolic Systems

One of the most developed areas of smooth ergodic theory is in the study of hyperbolic maps and attractors. This section defines hyperbolic maps and attractors, provides examples, and investigates their ergodic properties. See [10,11] and ► Hyperbolic Dynamical Systems for a thorough discussion of the topological and smooth properties of hyperbolic systems.

A *hyperbolic structure* on a compact f-invariant set $\Lambda \subset M$ is given by a Df-invariant splitting $T_\Lambda M = E^u \oplus E^s$ of the tangent bundle over Λ and constants $C, \mu > 1$ such that, for every $x \in \Lambda$ and $n \in \mathbb{N}$:

$$\begin{aligned} & v \in E^u(x) \implies \|D_x f^n(v)\| \geq C^{-1}\mu^n \|v\|\,, \\ \text{and} \quad & v \in E^s(x) \implies \|D_x f^n(v)\| \leq C\mu^{-n}\|v\|\,. \end{aligned}$$

A hyperbolic attractor for a map $f \colon M \to M$ is given by an open set $U \subset M$ such that: $f(U) \subset \overline{U}$, and such that the set $\Lambda = \bigcap_{n \geq 0} f^n(U)$ carries a hyperbolic structure. The set Λ is called the *attractor*, and U is an *attracting region*. A map $f \colon M \to M$ is *hyperbolic* if M decomposes (mod 0) into a finite union of attracting regions for hyperbolic attractors. Typically one assumes as well that the restriction of f to each attractor Λ_i is topologically transitive.

Every point p in a hyperbolic set Λ has smooth *stable manifold* $\mathcal{W}^{\mathrm{s}}(p)$ and *unstable manifold* $\mathcal{W}^{\mathrm{u}}(p)$, tangent, respectively, to the subspaces $E^s(p)$ and $E^{\mathrm{u}}(p)$. The set $\mathcal{W}^{\mathrm{s}}(p)$ is precisely the set of $q \in M$ such that $d(f^n(p), f^n(q))$ tends to 0 as $n \to \infty$, and it follows that $f(\mathcal{W}\mathrm{s}(p)) = \mathcal{W}^s(f(p))$. When f is a diffeomorphism, the unstable manifold $\mathcal{W}^{\mathrm{u}}(p)$ is uniquely defined and is the stable manifold of f^{-1}. When f is not invertible, local unstable manifolds exist, but generally are not unique. If Λ is a transitive hyperbolic attractor, then every unstable manifold of every point $p \in \Lambda$ is dense in Λ.

Examples of Hyperbolic Maps and Attractors

Expanding Maps The previous section mentioned briefly the C^r perturbations of the doubling map T_2. Such perturbations (as well as T_2 itself) are examples of *expanding maps*. A map $f\colon M \to M$ is *expanding* if there exist constants $\mu > 1$ and $C > 0$ such that, for every $x \in M$, and every nonzero vector $v \in T_x M$:

$$\|D_x f^n(v)\| \geq C\mu^n \|v\| ,$$

with respect to some (any) Riemannian metric on M. An expanding map is clearly hyperbolic, with $U = M$, E^s the trivial bundle, and $E^u = TM$. Any disk in M is a local unstable manifold for f.

Anosov Diffeomorphisms A diffeomorphism $f\colon M \to M$ is called *Anosov* if the tangent bundle splits as a direct sum $TM = E^u \oplus E^s$ of two Df-invariant subbundles, such that E^u is uniformly expanded and E^s is uniformly contracted by Df. Similarly, a flow $\varphi_t\colon M \to M$ is called Anosov if the tangent bundle splits as a direct sum $TM = E^u \oplus E^0 \oplus E^s$ of three $D\varphi_t$-invariant subbundles, such that E^0 is generated by $\dot{\varphi}$, E^u is uniformly expanded and E^s is uniformly contracted by $D\varphi_t$. Like expanding maps, an Anosov diffeomorphism is an Anosov attractor with $\Lambda = U = M$.

A simple example of a conservative Anosov diffeomorphism is a hyperbolic linear automorphism of the torus. Any matrix $A \in SL(n, \mathbb{Z})$ induces an automorphism of $\mathbb{R}^n$ preserving the integer lattice $\mathbb{Z}^n$, and so descends to an automorphism $f_A\colon T^n \to T^n$ of the n-torus $T^n = \mathbb{R}^n/\mathbb{Z}^n$. Since the determinant of A is 1, the diffeomorphism f_A preserves Lebesgue-Haar measure on T^n. In the case where none of the eigenvalues of A have modulus 1, the resulting diffeomorphism f_A is Anosov. The stablebundle E^s at $x \in T^n$ is the parallel translate to x of the sum of the contracted generalized eigenspaces of A, and the unstable bundle E^u at x is the translated sum of expanded eigenspaces.

In general, the invariant subbundles E^u and E^s of an Anosov diffeomorphism are integrable and tangent to a transverse pair of foliations $\mathcal{W}^u$ and $\mathcal{W}^s$, respectively (see, e. g [12]. for a proof of this). The leaves of $\mathcal{W}^s$ are uniformly contracted by f, and the leaves of $\mathcal{W}^u$ are uniformly contracted by f^{-1}. The leaves of these foliations are as smooth as f, but the tangent bundles to the leaves do not vary smoothly in the manifold. The regularity properties of these foliations play an important role in the ergodic properties of Anosov diffeomorphisms.

The first Anosov flows to be studied extensively were the geodesic flows for manifolds of negative sectional curvatures. As these flows are Hamiltonian, they are conservative. Eberhard Hopf showed in the 1930s that such geodesic flows for surfaces are ergodic with respect to Liouville measure [14]; it was not until the 1960s that ergodicity of all such flows was proved by Anosov [15]. The next section describes, in the context of Anosov diffeomorphisms, Hopf's method and important refinements due to Anosov and Sinai.

DA Attractors A simple way to produce a non-Anosov hyperbolic attractor on the torus is to start with an Anosov diffeomorphism, such as a linear hyperbolic automorphism, and deform it in a neighborhood of a fixed point, turning a saddle fixed point into a source, while preserving the stable foliation. If this procedure is carried out carefully enough, the resulting diffeomorphism is a dissipative hyperbolic diffeomorphism, called a *derived from Anosov (DA)* attractor. Other examples of hyperbolic attractors are the Plykin attractor and the solenoid. See [10].

Distortion Estimates

Before describing the ergodic properties of hyperbolic systems, it is useful to pause for a brief discussion of distortion estimates. Distortion estimates are behind almost every result in smooth ergodic theory. In the hyperbolic setting, distortion estimates are applied to the action of f on unstable manifolds to show that the volume distortion of f along unstable manifolds can be controlled for arbitrarily many iterates.

The example mentioned at the end of the previous section illustrates the ideas in a distortion estimate. Suppose that $f\colon S^1 \to S^1$ is a C^2 expanding map, such as a C^2 small perturbation of T_2. Then there exist constants $\mu > 1$ and $C > 0$ such that $(f^n)'(x) > C\mu^n$ for all x and n.

Let d be the degree of f. If I is a sufficiently small open interval in S^1, then for each n, $f^{-n}(I)$ is a union of d disjoint intervals. Furthermore, each of these intervals has diameter at most $C^{-1}\mu^{-n}$ times the diameter of I. It is now possible to justify the assertion in Fundamental Principle #3 in this context.

Lemma *There exists a constant $K \geq 1$ such that, for all $n \in \mathbb{N}$, and for all $x, y \in f^{-n}(I)$, one has:*

$$K^{-1} \leq \frac{(f^n)'(x)}{(f^n)'(y)} \leq K\,.$$

Proof Since f is C^2 and f' is bounded away from 0, the function $\alpha(x) = \log(f'(x))$ is C^1. In particular, α is Lipschitz continuous: there exists a constant $L > 0$ such that, for all $x, y \in S^1$, $|\alpha(x) - \alpha(y)| < Ld(x, y)$. For $n \geq 0$, let $\alpha_n(x) = \log((f^n)'(x))$. The Chain Rule implies that $\alpha_n(x) = \sum_{i=0}^{n-1} \alpha(f^i(x))$.

The expanding hypothesis on f implies that for all $x, y \in f^n(I)$ and for $i = 0, \ldots, n$, one has $d(f^i(x), f^i(y)) \leq C^{-1}\mu^{i-n} d(f^n(x), f^n(y)) \leq C^{-1}\mu^{i-n}$. Hence

$$\begin{aligned}
|\alpha_n(x) - \alpha_n(y)| &\leq \sum_{i=0}^{n-1} |\alpha(f^i(x)) - \alpha(f^i(y))| \\
&\leq L \sum_{i=0}^{n-1} d(f^i(x), f^i(y)) \\
&\leq L \sum_{i=0}^{n-1} C^{-1}\mu^{i-n} \\
&< LC^{-1}\mu^{-1}(1 - \mu^{-1})^{-1}\,.
\end{aligned}$$

Setting $K = \exp(LC^{-1}\mu^{-1}(1 - \mu^{-1})^{-1})$, one now sees that $(f^n)'(x)/(f^n)'(y)$ lies in the interval $[K^{-1}.K]$, proving the claim. □

In this distortion estimate, the function $\alpha\colon M \to \mathbb{R}$ is called a *cocycle*. The same argument applies to any Lipschitz continuous (or even Hölder continuous) cocycle.

Ergodicity of Expanding Maps

The ergodic properties of C^2 expanding maps are completely understood. In particular, every conservative expanding map is ergodic, and every expanding map is conservative. The proofs of these facts use Fundamental Principles #1 and 3 in a fairly direct way.

Every C^2 conservative expanding map is ergodic with respect to volume. The proof is a straightforward adaptation of the proof of Proposition 3 (see, e. g [2].). Here is a description of the proof for $M = S^1$. As remarked earlier, the proof of Proposition 3 adapts easily to a general expanding map $f\colon S^1 \to S^1$ once one shows that for every f-invariant set A, and every connected component J of $f^{-n}(S^1 \setminus \{p\})$, the quantity

$$\frac{\lambda(f^n(A \cap J)\colon f^n(J))}{\lambda(A\colon J)}$$

is bounded independently of n. This is a fairly direct consequence of the distortion estimate in Lemma 6.1 and is left as an exercise.

The same distortion estimates show that every C^2 expanding map is conservative, preserving a probability measure ν in the measure class of volume. Here is a sketch of the proof for the case $M = S^1$. To prove this, consider the push-forward $\lambda_n = f_*^n \lambda$. Then λ_n is equivalent to Lebesgue, and its Radon–Nikodym derivative $d\lambda_n \setminus d\lambda$ is the density function

$$\rho_n(x) = \sum_{y \in f^{-n}(x)} \frac{1}{\mathrm{jac}_y f^n}\,.$$

Since $f_*^n\lambda$ is a probability measure, it follows that $\int_{S^1} \rho_n \, \mathrm{d}\lambda = 1$. A simple argument using the distortion estimate above (and summing up over all d^n branches of f^{-n} at x) shows that there exists a constant $c \geq 1$ such that for all $x, y \in S^1$,

$$c^{-1} \leq \frac{\rho_n(x)}{\rho_n(y)} \leq c\,.$$

Since the integral of ρ_n is 1, the functions ρ_n are uniformly bounded away from 0 and ∞. It is easy to see that the measure $\nu_n = \frac{1}{n}\sum_{i=1}^{n} f_*^i\lambda$ has density $\frac{1}{n}\sum_{i=1}^{n} \rho_i$. Let ν be any subsequential weak* limit of ν_n; then ν is absolutely continuous, with density ρ bounded away from 0 and ∞. With a little more care, one can show that ρ is actually Lipschitz continuous.

As a passing comment, the ergodicity of ν and positivity of ρ imply that ν is the unique f-invariant measure absolutely continuous with respect to λ. With more work, one can show that ν is exact. See [2] for details.

Ergodicity of Conservative Anosov Diffeomorphisms

Like conservative C^2 expanding maps, conservative C^2 Anosov diffeomorphisms are ergodic. This subsection outlines a proof of this fact. Unlike expanding maps, however, Anosov diffeomorphisms need not be conservative. The subsection following this one describe a type of invariant measure that is "natural" with respect to volume, called a Sinai-Ruelle-Bowen (or SRB) measure. The central result for hyperbolic systems states that every hyperbolic attractor carries an SRB measure.

The Hopf Argument In the 1930s Hopf [14] proved that the geodesic flow for a compact, negatively-curved surface is ergodic. His method was to study the Birkhoff averages of continuous functions along leaves of the stable and unstable foliations of the flow. This type of argument has been used since then in increasingly general contexts, and has come to be known as the Hopf Argument.

The core of the Hopf Argument is very simple. To any $f\colon M \to M$ one can associate the *stable equivalence relation* $\sim_s$, where $x \sim_s y$ iff $\lim_{n\to\infty} d(f^n(x), f^n(y)) = 0$. Denote by $W^s(x)$ the stable equivalence class containing x. When f is invertible, one defines the *unstable equivalence relation* to be the stable equivalence relation for f^{-1}, and one denotes by $W^u(x)$ the unstable equivalence class containing x.

The first step in the Hopf Argument is to show that Birkhoff averages for continuous functions are constant along stable and unstable equivalence classes. Let $\phi\colon M \to \mathbb{R}$ be an integrable function, and let

$$\overline{\phi} = \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \phi \circ f^i \,. \tag{1}$$

Observe that if ϕ is continuous, then for every $x \in M$ and $x' \in W^s(x)$, $\lim_{n\to\infty} |\phi(f^n(x)) - \phi(f^i(x'))| = 0$. It follows immediately that $\overline{\phi}_f(x) = \overline{\phi}_f(x')$. In particular, if the limit in (1) exists at x, then it exists and is constant on $W^s(x)$.

Fundamental Principle #4: Birkhoff averages of continuous functions are constant along stable equivalence classes.

The next step of Hopf's argument confines itself to the situation where f is conservative and Anosov. In this case, f is invertible, the stable equivalence classes are precisely the leaves of the stable foliation $\mathcal{W}^s$, and the unstable equivalence classes are the leaves of the unstable foliation $\mathcal{W}^u$. Since f is conservative, the Ergodic Theorem implies that for every L^2 function ϕ, the function $\overline{\phi}_f$ is equal (mod 0) to the projection of ϕ onto the f-invariant functions in L^2. Since this projection is continuous, and the continuous functions are dense in L^2, to prove that f is ergodic, it suffices to show that the projection of any continuous function is trivial. That is, it suffices to show that for every continuous ϕ, the function $\overline{\phi}_f$ is constant (a.e.).

To this end, let $\phi\colon M \to \mathbb{R}$ be continuous. Since the f-invariant functions coincide with the f^{-1}-invariant functions, one obtains that $\overline{\phi}_f = \overline{\phi}_{f^{-1}}$ a.e. The previous argument shows $\overline{\phi}_f$ is constant along $\mathcal{W}^s$-leaves and $\overline{\phi}_{f^{-1}}$ is constant along $\mathcal{W}^u$-leaves. The desired conclusion is that $\overline{\phi}_f$ is a.e. constant. It suffices to show this in a local chart, since the manifold M is connected. In a local chart, after a smooth change of coordinates, one obtains a pair of transverse foliations $\mathcal{F}_1$, $\mathcal{F}_2$ of the cube $[-1,1]^n$ by disks, and a measurable function $\psi\colon [-1,1]^n \to \mathbb{R}$ that is constant along the leaves of $\mathcal{F}_1$ and constant along the leaves of $\mathcal{F}_2$.

When the foliations $\mathcal{F}_1$ and $\mathcal{F}_2$ are smooth (at least C^1), one can perform a further smooth change of coordinates so that $\mathcal{F}_1$ and $\mathcal{F}_2$ are transverse coordinate subspace foliations. In this case, Fubini's theorem implies that any measurable function that is constant along two transverse coordinate foliations is a.e. constant. This completes the proof in the case that the foliations $\mathcal{W}^s$ and $\mathcal{W}^u$ are smooth. In Hopf's original argument, the stable and unstable foliations were assumed to be C^1 foliations (a hypotheses satisfied in the examples he considered, due to low-dimensionality. See also [16], where a pinching condition on the curvature, rather than low dimensionality, implies this C^1 condition on the foliations.)

Absolute Continuity For a general Anosov diffeomorphism or flow, the stable and unstable foliations are not C^1, and so the final step in Hopf's orginal argument does not apply. The fundamental advance of Anosov and Anosov-Sinai was to prove that the stable and unstable foliations of an Anosov diffeomorphism (conservative or not) satisfy a weaker condition than smoothness, called *absolute continuity*. For conservative systems, absolute continuity is enough to finish Hopf's argument, proving that every C^2 conservative Anosov diffeomorphism is ergodic [15,17].

For a definition and careful discussion of absolute continuity of a foliation $\mathcal{F}$, see [13]. Two consequences of the absolute continuity of $\mathcal{F}$ are:

1. (AC1) If $A \subset M$ is any measurable set, then
$$\lambda(A) = 0 \iff \lambda_{\mathcal{F}(x)}(A) = 0\,, \qquad \text{for} \quad \lambda - \text{a.e. } x \in M\,,$$
where $\lambda_{\mathcal{F}(x)}$ denotes the induced Riemannian volume on the leaf of $\mathcal{F}$ through x.
2. (AC2) If τ is any small, smooth disk transverse to a local leaf of $\mathcal{F}$, and $T \subset \tau$ is a 0-set in τ (with respect to the induced Riemannian volume on τ), then the union of the $\mathcal{F}$ leaves through points in T has Lebesgue measure 0 in M.

The proof that $\mathcal{W}^s$ and $\mathcal{W}^u$ are absolutely continuous has a similar flavor to the proof that an expanding map has a unique absolutely continuous invariant measure (although the cocycles involved are Hölder continuous, rather than Lipschitz), and the facts are intimately related.

With absolute continuity of the stable and unstable foliations in hand, one can now prove:

Theorem 2 (Anosov) *Let f be a C^2, conservative Anosov diffeomorphism. Then f is ergodic.*

Proof By the Hopf Argument, it suffices to show that if ψ^s and ψ^u are L^2 functions with the following properties:

1. ψ^s is constant along leaves of $\mathcal{W}^s$,
2. ψ^u is constant along leaves of $\mathcal{W}^u$, and
3. $\psi^s = \psi^u$ a.e.,

then ψ^s (and so ψ^u as well) is constant a.e.

This is proved using the absolute continuity of $\mathcal{W}^u$ and $\mathcal{W}^s$. Since M is connected, one may argue this locally. Let G be the full measure set of $p \in M$ such that $\psi^s = \psi^u$. Absolute continuity of $\mathcal{W}^s$ (more precisely, consequence (AC1) of absolute continuity described above) implies that for almost every $p \in M$, G has full measure in $\mathcal{W}^s(p)$. Pick such a p. Then for almost every $q \in \mathcal{W}^s(p)$, $\psi^s(q) = \psi^u(p)$; defining G' to be the union over all $q \in \mathcal{W}^s(p) \cap G$ of $\mathcal{W}^u(q)$, one obtains that ψ^s is constant on $G \cap G'$. But now, since $\mathcal{W}^s(p) \cap G$ has full measure in $\mathcal{W}^s(p)$, the absolute continuity of $\mathcal{W}^u$ (consequence (AC2) above) implies that G' has full measure in a neighborhood of p. Hence ψ^s is a.e. constant in a neighborhood of p, completing the proof. □

SRB Measures

In the absence of a smooth invariant measure, it is still possible for a map to have an invariant measure that behaves naturally with respect to volume. In computer simulations one observes such measures when one picks a point x at random and plots many iterates of x; in many systems, the resulting picture is surprisingly insensitive to the initial choice of x. What appears to be happening in these systems is that the trajectory of almost every x in an open set U is converging to the support of a singular invariant probability measure μ. Furthermore, for any open set V, the proportion of forward iterates of x spent in V appears to converge to $\mu(V)$ as the number of iterates tends to ∞.

In the 1960s and 70s, Sinai, Ruelle and Bowen rigorously established the existence of these physically observable measures for hyperbolic attractors [18,19,20]. Such measures are now known as Sinai-Ruelle-Bowen (SRB) measures, and have been shown to exist for non-hyperbolic maps with some hyperbolic features. This subsection describes the construction of SRB measures for hyperbolic attractors.

An f-invariant probability measure μ is called an *SRB (or physical) measure* if there exists an open set $U \subset M$ containing the support of μ such that, for every continuous function $\phi\colon M \to \mathbb{R}$ and λ-a.e. $x \in U$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \phi(f^i(x)) = \int_M \phi \, d\mu \, .$$

The maximal open set U with this property is called the *basin* of f. To exclude the possibility that the SRB measure is supported on a periodic sink, one often adds the condition that at least one of the Lyapunov exponents of f with respect μ is positive. Other definitions of SRB measure have been proposed (see [21]). Note that every ergodic absolutely continuous invariant measure with positive density in an open set is an SRB measure. Note also that an SRB measure for f is not in general an SRB measure for f^{-1}, unless f preserves an ergodic absolutely continuous invariant measure.

Every transitive Anosov diffeomorphism carries a unique SRB measure. To prove this, one defines a sequence of probability measures ν_n on M as follows. Fix a point $p \in M$, and define ν_0 to be the normalized restriction of Riemannian volume to a ball B^u in $\mathcal{W}^u(p)$. Set $\nu_n = \frac{1}{n}\sum_{i=1}^{n} f_*^i \nu_0$. Distortion estimates show that the density of ν_n on its support inside $\mathcal{W}^u$ is bounded, above and below, independently of n. Passing to a subsequential weak* limit, one obtains a probability measure ν on M with bounded densities on $\mathcal{W}^u$-leaves.

To show that ν is an SRB measure, choose a point $q \in M$ in the support of ν. Since ν has positive density on unstable manifolds, almost every point in a neighborhood of q in $\mathcal{W}^u(q)$ is a regular point for f (that is, a point where the forward Birkhoff averages of every continuous function exist). A variation on the Hopf Argument, using the absolute continuity of $\mathcal{W}^s$, shows that ν is an ergodic SRB measure.

A similar argument shows that every transitive hyperbolic attractor admits an ergodic SRB measure. In fact this SRB measure has much stronger mixing properties, namely, it is Bernoulli. To prove this, one first constructs a Markov partition [22] conjugating the action of f to a Bernoulli shift. This map sends the SRB measure to a Gibbs state for a mixing Markov shift (see ► Pressure and Equilibrium States in Ergodic Theory). A result that subsumes all of the results in this section is:

Theorem 3 (Sinai, Ruelle, Bowen) *Let $\Lambda \subset M$ be a transitive hyperbolic attractor for a C^2 map $f\colon M \to M$. Then f has an ergodic SRB measure μ supported on Λ. Moreover: the disintegration of μ along unstable manifolds of Λ is equivalent to the induced Riemannian volume, the Lyapunov exponents of μ are all positive, and μ is Bernoulli.*

Beyond Uniform Hyperbolicity

The methods developed in the smooth ergodic theory of hyperbolic maps have been extended beyond the hyperbolic context. Two natural generalizations of hyperbolicity are:

- partial hyperbolicity, which requires uniform expansion of E^u and uniform contraction of E^s, but allows central directions at each point, in which the expansion and contraction is dominated by the behavior in the hyperbolic directions; and
- nonuniform hyperbolicity, which requires hyperbolicity along almost every orbit, but allows the expansion of E^u and the contraction of E^s to weaken near the exceptional set where there is no hyperbolicity.

This section discusses both generalizations.

Partial Hyperbolicity

Brin and Pesin [23] and independently Pugh and Shub [24] first examined the ergodic properties of partially hyperbolic systems soon after the work of Anosov and Sinai on hyperbolic systems. One says that a diffeomorphism $f\colon M \to M$ of a compact manifold M is *partially hyperbolic* if there is a nontrivial, continuous splitting of the tangent bundle, $TM = E^{\mathrm{s}} \oplus E^{\mathrm{c}} \oplus E^{\mathrm{u}}$, invariant under Df, such that E^{s} is uniformly contracted, E^{u} is uniformly expanded, and E^c is dominated, meaning that for some $n \geq 1$ and for all $x \in M$:

$$\|D_x f^n|_{E^s}\| < m(D_x f^n|_{E^c}) \leq \|D_x f^n|_{E^c}\| < m(D_x f^n|_{E^u}).$$

Partial hyperbolicity is a C^1-open condition: any diffeomorphism sufficiently C^1-close to a partially hyperbolic diffeomorphism is itself partially hyperbolic. For an extensive discussion of examples of partially hyperbolic dynamical systems, see the survey article [25] and the book [26]. Among these examples are: the time-1 map of an Anosov flow, the frame flow for a compact manifold of negative sectional curvature, and many affine transformations of compact homogeneous spaces. All of these examples preserve the volume induced by a Riemannian metric on M.

As in the Anosov case, the stable and unstable bundles E^s and E^u of a partially hyperbolic diffeomorphism are tangent to foliations, again denoted by $\mathcal{W}^s$ and $\mathcal{W}^u$ respectively [23]. Brin-Pesin and Pugh-Shub proved that these foliations are absolutely continuous.

A partially hyperbolic diffeomorphism $f\colon M \to M$ is *accessible* if any point in M can be reached from any other along an *su-path*, which is a concatenation of finitely many subpaths, each of which lies entirely in a single leaf of $\mathcal{W}^s$ or a single leaf of $\mathcal{W}^u$. Accessibility is a global, topological property of the foliations $\mathcal{W}^u$ and $\mathcal{W}^s$ that is the analogue of transversality of $\mathcal{W}^u$ and $\mathcal{W}^s$ for Anosov diffeomorphisms. In fact, the transversality of these foliations in the Anosov case immediately implies that every Anosov diffeomorphism is accessible. Fundamental Principle #4 suggests that accessibility might be related to ergodicity for conservative systems.

Conservative Partially Hyperbolic Diffeomorphisms Motivated by a breakthrough result with Grayson [27], Pugh and Shub conjectured that accessibility implies ergodicity, for a C^2, partially hyperbolic conservative diffeomorphism [28]. This conjecture has been proved under the hypothesis of center bunching [29], which is a mild spectral condition on the restriction of Df to the center bundle E^{c}. Center bunching is satisfied by most examples of interest, including all partially hyperbolic diffeomorphisms with $\dim(E^c) = 1$. The proof in [29] is a modification of the Hopf Argument using Lebesgue density points and a delicate analysis of the geometric and measure-theoretic properties of the stable and unstable foliations.

In the same article, Pugh and Shub also conjectured that accessibility is a widespread phenomenon, holding for an open and dense set (in the C^{r} topology) of partially hyperbolic diffeomorphisms. This conjecture has been proved completely for $r = 1$ [30], and for all r, with the additional assumption that the central bundle E^{c} is one dimensional [31].

Together, these two conjectures imply the third, central conjecture: in [28]:

Conjecture 1 (Pugh–Shub) *For any $r \geq 2$, the C^r, conservative partially hyperbolic diffeomorphisms contain a C^r open and dense set of ergodic diffeomorphisms.*

The validity of this conjecture in the absence of center bunching is currently an open question.

Dissipative Partially Hyperbolic Diffeomorphisms There has been some progress in constructing SRB-type measures for dissipative partially hyperbolic diffeomorphisms, but the theory is less developed than in the conservative case. Using the same construction as for Anosov diffeomorphisms, one can construct invariant probability measures that are smooth along the $\mathcal{W}^u$ foliation [32]. Such measures are referred to as u-Gibbs measures. Since the stable bundle E^{s} is not transverse to the unstable bundle E^{u}, the Anosov argument cannot be carried through to show that u-Gibbs measures are SRB measures.

Nonetheless, there are conditions that imply that a u-Gibbs measure is an SRB measure: for example, a u-Gibbs measure is SRB if it is the unique u-Gibbs measure [33], if the bundle $E^s \oplus E^c$ is nonuniformly contracted [34], or if the bundle $E^u \oplus E^c$ is nonuniformly expanded [35]. The proofs of the latter two results use Pesin Theory, which is explained in the next subsection.

SRB measures have also been constructed in systems where E^c is nonuniformly hyperbolic [36], and in (noninvertible) partially hyperbolic covering maps where E^c is 1-dimensional [37]. It is not known whether accessibility plays a role in the existence of SRB measures for dissipative, non-Anosov partially hyperbolic diffeomorphisms.

Nonuniform Hyperbolicity

The concept of Lyapunov exponents gives a natural way to extend the notion of hyperbolicity to systems that behave hyperbolically, but in a nonuniform manner. The fundamental principles described above, suitably modified, apply to these nonuniformly hyperbolic systems and allow for the development of a smooth ergodic theory for these systems. This program was initially proposed and carried out by Yakov Pesin in the 1970s [38] and has come to be known as *Pesin theory*.

Oseledec's Theorem implies that if a smooth map f satisfying the condition $m(D_x f) > 0$ preserves a probability measure ν, then for ν-a.e. $x \in M$ and every nonzero vector $v \in T_x M$, the limit

$$\lambda(x, v) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \log \| D_x f^i(v) \|$$

exists. The number $\lambda(x, v)$ is called the *Lyapunov exponent at x in the direction of v*. For each such x, there are finitely many possible values for the exponent $\lambda(x, v)$, and the function $x \mapsto \lambda(x, \cdot)$ is measurable. See the discussion of Oseledec's Theorem in ▶ Ergodic Theorems.

Let f be a smooth map. An f-invariant probability measure μ is *hyperbolic* if the Lyapunov exponents of μ-a.e. point are all nonzero. Observe that any invariant measure of a hyperbolic map is a hyperbolic measure.

A conservative diffeomorphism $f: M \to M$ is *nonuniformly hyperbolic* if the invariant measure equivalent to volume is hyperbolic. The term "nonuniform" is a bit misleading, as uniformly hyperbolic conservative systems are also nonuniformly hyperbolic. Unlike uniform hyperbolicity, however, nonuniform hyperbolicity allows for the *possibility* of different strengths of hyperbolicity along different orbits.

Nonuniformly hyperbolic diffeomorphisms exist on all manifolds [39,40], and there are C^1-open sets of nonuniformly hyperbolic diffeomorphisms that are not Anosov diffeomorphisms [41]. In general, it is a very difficult problem to establish whether a given map carries a hyperbolic measure that is nonsingular with respect to volume.

Hyperbolic Blocks As mentioned above, the derivative of f along almost every orbit of a nonuniformly hyperbolic system looks like the derivative down the orbit of a uniformly hyperbolic system; the nonuniformity can be detected only by examining a positive measure set of orbits. Recall that Lusin's Theorem in measure theory states that every Borel measurable function is continuous on the complement of an arbitrarily small measure set. A sort of analogue of Lusin's theorem holds for nonuniformly hyperbolic maps: every C^2, nonuniformly hyperbolic diffeomorphism is uniformly hyperbolic on a (noninvariant) compact set whose complement has arbitrarily small measure. The precise formulation of this statement is omitted, but here are some of its salient features.

If μ is a hyperbolic measure for a C^2 diffeomorphism, then attached to μ-a.e. point $x \in M$ are transverse, smooth stable and unstable manifolds for f. The collection of all stable manifolds is called the *stable lamination* for f, and the collection of all unstable manifolds is called the *unstable lamination* for f. The stable lamination is invariant under f, meaning that f sends the stable manifold at x into the stable manifold for $f(x)$. The stable manifold through x is contracted uniformly by all positive iterates of f in a neighborhood of x. Analogous statements hold for the unstable manifold of x, with f replaced by f^{-1}.

The following quantites vary measurably in $x \in M$:

- the (inner) radii of the stable and unstable manifolds through x,
- the angle between stable and unstable manifolds at x, and
- the rates of contraction in these manifolds.

The stable and unstable laminations of a nonuniformly hyperbolic system are absolutely continuous. The precise definition of absolute continuity here is slightly different than in the uniformly and partially hyperbolic setting, but the consequences (AC1) and (AC2) of absolute continuity continue to hold.

Ergodic Properties of Nonuniformly Hyperbolic Diffeomorphisms Since the stable and unstable laminations are absolutely continuous, the Hopf Argument can be applied in this setting to show:

Theorem 4 (Pesin) *Let f be C^2, conservative and nonuniformly hyperbolic. Then there exists a (mod 0) partition $\mathcal{P}$ of*

M into countably many f-invariant sets of positive volume such that the restriction of f to each $P \in \mathcal{P}$ is ergodic.

The proof of this theorem is also exposited in [42]. The countable partition can in examples be countably infinite; nonuniform hyperbolicity alone does not imply ergodicity.

The Dissipative Case As mentioned above, establishing the existence of a nonsingular hyperbolic measure is a difficult problem in general. In systems with some global form of hyperbolicity, such as partial hyperbolicity, it is sometimes possible to "borrow" the expansion from the unstable direction and lend it to the central direction, via a small perturbation. Nonuniformly hyperbolic attractors have been constructed in this way [43]. This method is also behind the construction of a C^1 open set of nonuniformly hyperbolic diffeomorphisms in [41].

For a given system of interest, it is sometimes possible to prove that a given invariant measure is hyperbolic by establishing an *approximate* form of hyperbolicity. The idea, due to Wojtkowski and called the *cone method*, is to isolate a measurable bundle of cones in TM defined over the support of the measure, such that the cone at a point x is mapped by $D_x f$ into the cone at $f(x)$. Intersecting the images of these cones under all iterates of Df, one obtains an invariant subbundle of TM over the support of f that is nonuniformly expanded.

Lai-Sang Young has developed a very general method [44] for proving the existence of SRB measures with strong mixing properties in systems that display "some hyperbolicity". The idea is to isolate a region X in the manifold where the first return map is hyperbolic and distortion estimates hold. If this can be done, then the map carries a mixing, hyperbolic SRB measure. The precise rate of mixing is determined by the properties of the return-time function to X; the longer the return times, the slower the rate of mixing.

More results on the existence of hyperbolic measures are discussed in the next section.

An important subject in smooth ergodic theory is the relationship between entropy, Lyapunov exponents, and dimension of invariant measures of a smooth map. Significant results in this area include the Pesin entropy formula [45], the Ruelle entropy inequality [46], the entropy-exponents-dimension formula of Ledrappier–Young [47,48], and the proof by Barreira-Pesin-Schmeling that hyperbolic measures have a well-defined dimension [49]. ► Hyperbolic Dynamical Systems contains a discussion of these results; see this entry there for further information.

The Presence of Critical Points and Other Singularities

Now for a discussion of the aforementioned technical difficulties that arise in the presence of singularities and critical points for the derivative.

Singularities, that is, points where Df (or even f) fails to be defined, arise naturally in the study of billiards and hard sphere gases. The first subsection discusses some progress made in smooth ergodic theory in the presence of singularities.

Critical points, that is, points where Df fails to be invertible, appear inescapably in the study of noninvertible maps. This type of complication already shows up for noninvertible maps in dimension 1, in the study of unimodal maps of the interval. The second subsection discusses the technique of parameter exclusion, developed by Jakobson, which allows for an ergodic analysis of a parametrized family of maps with criticalities.

The technical advances used to overcome these issues in the interval have turned out to have applications to dissipative, nonhyperbolic, diffeomorphisms in higher dimension, where the derivative is "nearly critical" in places. The last subsection describes extensions of the parameter exclusion technique to these near-critical maps.

Hyperbolic Billiards and Hard Sphere Gases

In the 1870s the physicist Ludwig Boltzmann hypothesized that in a mechanical system with many interacting particles, physical measurements (observables), averaged over time, will converge to their expected value as time approaches infinity. The underlying dynamical system in this statement is a Hamiltonian system with many degrees of freedom, and the "expected value" is with respect to Liouville measure. Loosely phrased in modern terms, Boltzmann's hypothesis states that a generic Hamiltonian system of this form will be ergodic on constant energy submanifolds. Reasoning that the time scales involved in measurement of an observable in such a system are much larger than the rate of evolution of the system, Boltzmann's hypothesis allowed him to assume that physical quantities associated to such a system behave like constants.

In 1963, Sinai revived and formalized this ergodic hypothesis, stating it in a concrete formulation known as the Boltzmann-Sinai Ergodic Hypothesis. In Sinai's formulation, the particles were replaced by N hard, elastic spheres, and to compactify the problem, he situated the spheres on a k-torus, $k = 2, 3$. The Boltzmann-Sinai Ergodic Hypothesis is the conjecture that the induced Hamiltonian system on the $2kN$-dimensional configuration space is ergodic on constant energy manifolds, for any $N \geq 2$.

Sinai verified this conjecture for $N = 2$ by reducing the problem to a billiard map in the plane. As background for Sinai's result, a brief discussion of planar billiard maps follows.

Let $D \subset \mathbb{R}^s$ be a connected region whose boundary ∂D is a collection of closed, piecewise smooth simple curves the plane. The *billiard map* is a map defined (almost everywhere) on $\partial D \times [-\pi, \pi]$. To define this map, one identifies each point $(x, \theta) \in \partial D \times [-\pi, \pi]$ with an inward-pointing tangent vector at x in the plane, so that the normal vector to ∂D at x corresponds to the pair $(x, \pi/2)$. This can be done in a unique way on the smooth components of ∂D. Then $f(x, \theta)$ is obtained by following the ray originating at (x, θ) until it strikes the boundary ∂D for the first time at (x', θ'). Reflecting this vector about the normal at x', define $f(x, \theta) = (x', \pi - \theta')$.

It is not hard to see that the billiard map is conservative. The billiard map is piecewise smooth, but not in general smooth: the degree of smoothness of f is one less than the degree of smoothness of ∂D. In addition to singularities arising from the corners of the table, there are singularities arising in the *second* derivative of f at the tangent vectors to the boundary.

In the billiards studied studied by Sinai, the boundary ∂D consists of a union of concave circular arcs and straight line segments. Similar billiards, but with convex circular arcs, were first studied by Bunimovich [51]. Sinai and Bunimovich proved that these billiards are ergodic and nonuniformly hyperbolic. For the Boltzmann-Sinai problem with $N \geq 3$, the relevant associated dynamical system is a higher dimensional billiard table in euclidean space, with circular arcs replaced by cylindrical boundary components.

In a planar billiard table with circular/flat boundary, the behavior of vectors encountering a flat segment of boundary is easily understood, as is the behavior of vectors meeting a circular segment in a neighborhood of the normal vector. If the billiard map is ergodic, however, every open set of vectors will meet the singularities in the table infinitely many times. To establish the nonuniform hyperbolicity of such billiard tables via conefieds, it is therefore necessary to understand precisely the fraction of time orbits spend near these singularities. Furthermore, to use the Hopf argument to establish ergodicity, one must avoid the singularities in the second derivative, where distortion estimates break down. The techniques for overcoming these obstacles involve imposing restrictions on the geometry of the table (even more so for higher dimensional tables), and are well beyond the scope of this paper.

The study of hyperbolic billiards and hard sphere gases has a long and involved history. See the articles [50] and [52] for a survey of some of the results and techniques in the area. A discussion of methods in singular smooth ergodic theory, with particular applications to the Lorentz attractor, can be found in [53]. Another, more classical, reference is [54], which contains a formulation of properties on a critical set, due to Katok–Strelcyn, that are useful in establishing ergodicity of systems with singularities.

Interval Maps and Parameter Exclusion

The logistic family of maps $f_t\colon x \mapsto tx(1-x)$ defined on the interval $[0, 1]$ is very simple to define but exhibits an astonishing variety of dynamical features as the parameter t varies. For small positive values of t, almost every point in I is attracted under the map f_t to the sink at -1. For values of $t > 4$, the map has a repelling hyperbolic Cantor set. As the value of t increases between 0 and 4, the map f_t undergoes a cascade of period-doubling bifurcations, in which a periodic sink of period 2^n becomes repelling and a new sink of period 2^{n+1} is born. At the accumulation of period doubling at $t \approx 3.57$, a periodic point of period 3 appears, forcing the existence of periodic points of all periods. The dynamics of f_t for t close to 4 has been the subject of intense inquiry in the last 20 years.

The map f_t, for t close to 4, shares some of the features of the doubling map T_2; it is 2-to-1, except at the critical point $\frac{1}{2}$, and it is uniformly expanding in the complement of a neighborhood of this critical point. Because this neighborhood of the critical point is not invariant, however, the only invariant sets on which f_t is uniformly hyperbolic have measure zero. Furthermore, the second derivative of f_t vanishes at the critical point, making it impossible to control distortion for orbits that spend too much time near the critical point.

Despite these serious obstacles, Michael Jakobson [55] found a method for constructing absolutely continuous invariant measures for maps in the logistic family. The method has come to be known as *parameter exclusion* and has seen application far beyond the logistic family. As with billiards, it is possible to formulate geometric conditions on the map f_t that control both expansion (hyperbolicity) and distortion on a positive measure set. As these conditions involve understanding infinitely many iterates of f_t, they are impossible to verify for a given parameter value t.

Using an inductive formulation of this condition, Jakobson showed that the set of parameters t near 4 that fail to satisfy the condition at iterate n have exponentially small measure (in n). He thereby showed that for a positive Lebesgue measure set of parameter values t, the map f_t has an absolutely continuous invariant measure [55]. This measure is ergodic (mixing) and has a positive Lyapunov exponent. The delicacy of Jakobson's approach is confirmed by the fact that for an open and dense set of parameter values, almost every orbit is attracted to a periodic sink, and so f_t has no absolutely continuous invariant measure [56,57]. Jakobson's method applies not only to the logistic family but to a very general class of C^3 one-parameter families of maps on the interval.

Near-Critical Diffeomorphisms

Jakobson's method in one dimension proved to extend to certain highly dissipative diffeomorphisms. The seminal paper in this extension is due to Benedicks and Carleson; the method has since been extended in a series of papers [59,60,61] and has been formulated in an abstract setting [62].

This extension turns out to be highly nontrivial, but it is possible to describe informally the similarities between the logistic family and higher-dimensional "near critical" diffeomorphisms. The diffeomorphisms to which this method applies are crudely hyperbolic with a one dimensional unstable direction. Roughly this means that in some invariant region of the manifold, the image of a small ball under f will be stretched significantly in one direction and shrunk in all other directions. The directions of stretching and contraction are transverse in a large proportion of the invariant region, but there are isolated "near critical" subregions where expanding and contracting directions are nearly tangent.

The dynamics of such a diffeomorphism are very close to 1-dimensional if the contraction is strong enough, and the diffeomorphism resembles an interval map with isolated critical points, the critical points corresponding to the critical regions where stable and unstable directions are tangent.

An illustration of this type of dynamics is the Hénon family of maps $f_{a,b}\colon (x,y) \mapsto (1 - ax^2 + by, x)$, the original object of study in Benedicks-Carlesson's work. When the parameter b is set to 0, the map $f_{a,b}$ is no longer a diffeomorphism, and indeed is precisely a projection composed with the logistic map. For small values of b and appropriate values of a, the Hénon map is strongly dissipative and displays the near critical behavior described in the previous paragraph. In analogy to Jakobson's result, there is a positive measure set of parameters near $b = 0$ where $f_{a,b}$ has a mixing, hyperbolic SRB measure.

See [63] for a detailed exposition of the parameter exclusion method for Hénon-like maps.

Future Directions

In addition to the open problems discussed in the previous sections, there are several general questions and problems worth mentioning:

- What can be said about systems with everywhere vanishing Lyapunov exponents? Open sets of such systems exist in arbitrary dimension. Pesin theory carries into the nonuniformly hyperbolic setting the basic principles from uniformly hyperbolic theory (in particular, Fundamental Principles #3 and 4 above). To what extent do properties of isometric and unipotent systems (for example, Fundamental Principle #2) extend to conservative systems all of whose Lyaponov exponents vanish?
- Can one establish the existence of and analyze in general the conservative systems on surfaces that have two positive measure regimes: one where Lyapunov exponents vanish, and the other where they are nonzero? Such systems are conjectured exist in the presence of KAM phenomena surrounding elliptic periodic points.
- On a related note, how common are conservative systems whose Lyapunov exponents are nonvanishing on a positive measure set? See [64] for a discussion.
- Find a broad description of those dissipative systems that admit finitely (or countably) many physical measures. Are such systems dense among all dissipative systems, or possibly generic among a restricted class of systems? See [41,65] for several questions and conjectures related to this problem.
- Extend the methods in the study of systems with singularities to other specific systems of interest, including the infinite dimensional systems that arise in the study of partial differential equations.
- Carry the methods of smooth ergodic theory further into the study of smooth actions of discrete groups (other than the integers) on manifolds. When do such actions admit (possibly non-invariant) "physical" measures?

There are other interesting open areas of future inquiry, but this gives a good sample of the range of possibilities.

Bibliography

1. Arnol'd VI, Avez A (1986) Ergodic problems of classical mechanics. Translated from the French by Avez, A. Benjamin WA, Inc., New York Amsterdam
2. Mañé R (1987) Ergodic theory and differentiable dynamics. Translated from the Portuguese by Silvio L. Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)], vol 8. Springer, Berlin

3. Baladi V (2000) Positive transfer operators and decay of correlations. Advanced Series in Nonlinear Dynamics, 16. World Scientific Publishing Co., Inc., River Edge
4. Kifer Y (1986) Ergodic theory of random transformations. Progress in Probability and Statistics, vol 10. Birkhäuser, Boston
5. Kifer Y (1988) Random perturbations of dynamical systems. Progress in Probability and Statistics, 16. Birkhäuser, Boston
6. Liu PD, Qian M (1995) Smooth ergodic theory of random dynamical systems. Lecture Notes in Mathematics, vol 1606. Springer, Berlin
7. Bonatti C, Dìaz LJ, Viana M (2005) Dynamics beyond uniform hyperbolicity. A global geometric and probabilistic perspective. Encyclopaedia of Mathematical Sciences, 102. Mathematical Physics, III. Springer, Berlin
8. Cornfeld IP, Fomin SV, Sinaĭ YG (1982) Ergodic theory. Translated from the Russian by A. B. Sosinskiĭ. Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol 245. Springer, New York
9. Hirsch MW (1979) Differential topology. Graduate Texts in Mathematics, No. 33. Springer, New York Heidelberg
10. Robinson C (1995) Dynamical systems. Stability, symbolic dynamics, and chaos. Studies in Advanced Mathematics. CRC Press, Boca Raton
11. Katok A, Hasselblatt B (1995) Introduction to the modern theory of dynamical systems. With a supplementary chapter by Katok and Leonardo Mendoza. Encyclopedia of Mathematics and its Applications, vol 54. Cambridge University Press, Cambridge
12. Hirsch MW, Pugh CC, Shub M (1977) Invariant manifolds. Lecture Notes in Mathematics, vol 583. Springer, Berlin New York
13. Brin M, Stuck G (2002) Introduction to dynamical systems. Cambridge University Press, Cambridge
14. Hopf E (1939) Statistik der geodätischen Linien in Mannigfaltigkeiten negativer Krümmung. Ber Verh Sachs Akad Wiss Leipzig 91:261–304 (German)
15. Anosov DV (1967) Geodesic flows on closed Riemann manifolds with negative curvature. Proceedings of the Steklov Institute of Mathematics, No. 90 1967. Translated from the Russian by S. Feder. American Mathematical Society, Providence
16. Sinaĭ JG (1961) Geodesic flows on compact surfaces of negative curvature. Dokl Akad Nauk SSSR 136:549–552 (Russian); translated as Soviet Math Dokl 2:106–109
17. Anosov DV, Sinaĭ JG (1967) Certain smooth ergodic systems. (Russian) Uspehi Mat Nauk 22 no. 5(137):107–172
18. Sinaĭ JG (1972) Gibbs measures in ergodic theory. (Russian) Uspehi Mat Nauk 27 no. 4(166):21–64
19. Ruelle D (1976) A measure associated with axiom-A attractors. Amer J Math 98(3):619–654
20. Bowen R (1975) Equilibrium states and the ergodic theory of Anosov diffeomorphisms. Lecture Notes in Mathematics, vol 470. Springer, Berlin New York
21. Young LS (2002) What are SRB measures, and which dynamical systems have them? Dedicated to David Ruelle and Yasha Sinai on the occasion of their 65th birthdays. J Statist Phys 108(5–6):733–754
22. Bowen R (1970) Markov partitions for Axiom A diffeomorphisms. Amer J Math 92:725–747
23. Brin MI, Pesin JB (1974) Partially hyperbolic dynamical systems. Izv Akad Nauk SSSR Ser Mat 38:170–212 (Russian)
24. Pugh C, Shub M (1972) Ergodicity of Anosov actions. Invent Math 15:1–23
25. Burns K, Pugh C, Shub M, Wilkinson A (2001) Recent results about stable ergodicity. Smooth ergodic theory and its applications. (Seattle, 1999), 327–366. Proc Sympos Pure Math, 69, Amer Math Soc, Providence
26. Pesin YB (2004) Lectures on partial hyperbolicity and stable ergodicity. Zurich Lectures in Advanced Mathematics. European Mathematical Society (EMS), Zürich
27. Grayson M, Pugh C, Shub M (1994) Stably ergodic diffeomorphisms. Ann of Math 140(2):295–329
28. Pugh C, Shub M (1996) Stable ergodicity and partial hyperbolicity. International Conference on Dynamical Systems (Montevideo, 1995), 182–187, Pitman Res Notes Math Ser, 362, Longman, Harlow
29. Burns K, Wilkinson A, On the ergodicity of partially hyperbolic systems. Ann of Math. To appear
30. Dolgopyat D, Wilkinson A (2003) Stable accessibility is C^1 dense. Geometric methods in dynamics. II. Astérisque No. 287
31. Rodríguez HA, Rodríguez HF, Ures R (2008) Partially hyperbolic systems with 1D-center bundle. Invent Math 172(2)
32. Pesin YB, Sinaĭ YG (1983) Gibbs measures for partially hyperbolic attractors. Ergod Theor Dynam Syst 2(3–4):417–438
33. Dolgopyat D (2004) On differentiability of SRB states for partially hyperbolic systems. Invent Math 155(2):389–449
34. Bonatti C, Viana M (2000) SRB measures for partially hyperbolic systems whose central direction is mostly contracting. Israel J Math 115:157–193
35. Alves JF, Bonatti C, Viana M (2000) SRB measures for partially hyperbolic systems whose central direction is mostly expanding. Invent Math 140(2):351–398
36. Burns K, Dolgopyat D, Pesin Y, Pollicott Mark Stable ergodicity for partially hyperbolic attractors with negative central exponents. Preprint
37. Tsujii M (2005) Physical measures for partially hyperbolic surface endomorphisms. Acta Math 194(1):37–132
38. Pesin JB (1977) Characteristic Ljapunov exponents, and smooth ergodic theory. Uspehi Mat Nauk 32(196):55–112, 287 (Russian)
39. Katok A (1979) Bernoulli diffeomorphisms on surfaces. Ann Math 110(2):529–547
40. Dolgopyat D, Pesin Y (2002) Every compact manifold carries a completely hyperbolic diffeomorphism. Ergod Theor Dynam Syst 22(2):409–435
41. Shub M, Wilkinson A (2000) Pathological foliations and removable zero exponents. Invent Math 139(3):495–508
42. Pugh C, Shub M (1989) Ergodic attractors. Trans Amer Math Soc 312(1):1–54
43. Viana M (1997) Multidimensional nonhyperbolic attractors. Inst Hautes Études Sci Publ Math No 85:63–96
44. Young LS (1998) Statistical properties of dynamical systems with some hyperbolicity. Ann Math 147(2):585–650
45. Pesin JB (1976) Characteristic Ljapunov exponents, and ergodic properties of smooth dynamical systems with invariant measure. Dokl Akad Nauk SSSR 226 (Russian)
46. Ruelle D (1978) An inequality for the entropy of differentiable maps. Bol Soc Brasil Mat 9(1):83–87
47. Ledrappier F, Young LS (1985) The metric entropy of diffeomorphisms. I. Characterization of measures satisfying Pesin's entropy formula. Ann Math 122(2):509–539
48. Ledrappier F, Young LS (1985) The metric entropy of diffeomorphisms. II. Relations between entropy, exponents and dimension. Ann Math 122(2):540–574

49. Barreira L, Pesin Y, Schmeling J (1999) Dimension and product structure of hyperbolic measures. Ann Math 149(2):755–783
50. Szász D (2000) Boltzmann's ergodic hypothesis, a conjecture for centuries? Hard ball systems and the Lorentz gas, Encycl Math Sci, vol 101. Springer, Berlin, pp 421–448
51. Bunimovič LA (1974) The ergodic properties of certain billiards. (Russian) Funkcional. Anal i Priložen 8(3):73–74
52. Chernov N, Markarian R (2003) Introduction to the ergodic theory of chaotic billiards. Second edition. Publicações Matemáticas do IMPA. IMPA Mathematical Publications 24° Colq́uio Brasileiro de Matemática. 26th Brazilian Mathematics Colloquium Instituto de Matemática Pura e Aplicada (IMPA), Rio de Janeiro
53. Araújo V, Pacifico MJ (2007) Three Dimensional Flows. Publicações Matemáticas do IMPA. [IMPA Mathematical Publications] 26° Colq́uioquio Brasileiro de Matemática. [26th Brazilian Mathematics Colloquium] Instituto de Matemática Pura e Aplicada (IMPA), Rio de Janeiro
54. Katok A, Strelcyn JM, Ledrappier F, Przytycki F (1986) Invariant manifolds, entropy and billiards; smooth maps with singularities. Lecture Notes in Mathematics, 1222. Springer, Berlin
55. Jakobson MV (1981) Absolutely continuous invariant measures for one-parameter families of one-dimensional maps. Comm Math Phys 81(1):39–88
56. Graczyk J, Światek G (1997) Generic hyperbolicity in the logistic family. Ann Math 146(2):1–52
57. Lyubich M (1999) Feigenbaum–Coullet–Tresser universality and Milnor's hairiness conjecture. Ann Math 149(2):319–420
58. Benedicks M, Carleson L (1991) The dynamics of the Hénon map. Ann Math 133(2):73–169
59. Mora L. Viana M (1993) Abundance of strange attractors. Acta Math 171(1):1–71
60. Benedicks M, Young LS (1993) Sinaĭ–Bowen–Ruelle measures for certain Hénon maps. Invent Math 112(3):541–576
61. Benedicks M, Viana M (2001) Solution of the basin problem for Hénon-like attractors. Invent Math 143(2):375–434
62. Wang QD, Young LS (2008) Toward a theory of rank one attractors. Ann Math 167(2)
63. Viana M, Lutstsatto S (2003) Exclusions of parameter values in Hénon-type systems. (Russian) Uspekhi Mat Nauk 58 (2003), no. 6(354), 3–44; translation in Russian Math. Surveys 58(6):1053–1092
64. Bochi J, Viana M (2004) Lyapunov exponents: how frequently are dynamical systems hyperbolic? Modern dynamical systems and applications. Cambridge Univ Press, Cambridge, pp 271–297
65. Palis J (2005) A global perspective for non-conservative dynamics. Ann Inst H Poincaré Anal Non Linéaire 22(4):485–507

Social Cognitive Complexity

JÜRGEN KLÜVER
Duisburg-Essen University, Essen, Germany

Article Outline

Glossary

Attractor A sub space of the state space of a certain system that the system does not leave anymore. If the sub space consists of only one point it is a point attractor.

Dialectical relation A relation between two persons or generally two sides representing a tension and a mutual dependency. In more contemporary terms a dialectical relation is a certain form of feed back relation.

Homo oeconomicus An actor who acts according to the assumptions of rational choice.

Homo sociologicus The assumption that man is basically a being following social rules in contrast to *homo oeconomicus*.

Meta rules Rules that operate on rules of interaction and change these rules.

Rational choice The theoretical assumption that man is a rational and egoistical actor, who tries to maximize his profit and selects in a certain situation the best strategies.

Socialization The biographical process by which an individual learns and internalizes the cultural norms and social rules of his specific society. The result is a social actor.

Social-cognitive systems Complex dynamical systems that consist of social actors and contain at least two different levels, namely a social level and a cognitive one.

Socio-cultural evolution The development of societies in the mutually interdependent dimensions of social structure and culture.

Universal modeling schema A schema that consists of different interdependent levels and each level is modeled as a complex dynamical system.

Definition of the Subject

Social-cognitive systems certainly belong to the most complex systems that we know. The reason for this complexity is due to the fact that social-cognitive systems at least consist of two different levels, namely a social and a cognitive one that permanently interact and mutually influence each other. In terms of the social and cognitive sciences, social-cognitive systems consist of social actors who interact and whose actions are determined by certain social rules. Yet these actors are also complex dynamical systems, i. e., cognitive ones. The cognitive processes of the individual actors also influence the specific actions. The social rules determine in some part the cognitive processes that in turn generate the according actions. But also in turn the cognitive processes may determine the actions in such a way that the social rules are influenced and even changed. That is for example the case in times of revolutions and/or political reforms. In this sense social-cognitive systems are determined by a permanent interdependent dynamics between these levels and that is why a modeling and precise analysis of these systems is rather difficult. In addition, social-cognitive systems are able to adapt to changing environmental conditions. This means that these systems are able to change their rules and structure.

Introduction

Social-cognitive complexity, although of course not in this term, has been the subject of social thinking since the beginning of reflections on society and social actions. Already one of the earliest works on society in human history, namely Plato's *Politeia* (The Republic), emphasizes the importance of the interdependency of mind and social structure: A certain desired social structure, i. e. the differentiation of society into different social classes, can be generated and reproduced only by influencing the minds of the social actors via education. A suited education cares for the adequate thinking of the individuals, which in turn guarantees the production and reproduction of the class structure by the social actions of the educated actors. The class structure finally determines the minds of the individuals and produces the wished forms of socially determined thinking. Plato's ideal society is certainly not a paradigm for democratic societies but the logic of Plato's arguments already demonstrates the necessity to take account of both social and cognitive levels if one wants to understand society and the socially determined actions of human beings. It is not by chance that all authors who thought about utopian societies also stressed the importance of education, i. e. the influencing of minds to produce and reproduce social structures, for example Huxley and Orwell.

More than two millennia later, Karl Marx in his theory of Historical Materialism [27] dealt in a similar way with the subject of social-cognitive complexity. He postulated a "dialectical" relation between consciousness or mind respectively and social structure – the social conditions (*soziale Verhältnisse*) in his terms. The social conditions determine the minds of social actors and produce that way an ideological consciousness; accordingly the social actors reproduce the social conditions by their ideologically determined actions. That is the essence of Marx's famous base-superstructure theorem: the social conditions determine the consciousness, which in turn produces the ideological parts of society (In the German original: *Das gesellschaftliche oder materielle Sein bestimmt das Bewusstsein*). Yet in times of societal crisis the consciousness of certain social actors, namely the members of revolutionary classes, the consciousness is able to overcome the ideological constraints and produces by the social actions of revolution a new social structure. In such times historical progress is generated by the determination of social structures by the social minds of the revolutionary actors. That is why Marx called the relationship between mind and social structure a dialectical one: both sides of the relationship form a dynamical tension, influence the other and are influenced in turn by the other. If one substitutes the philosophical term of dialectic by the more precise contemporary term of feed back, one may say that Marx postulated social-cognitive systems as systems whose dynamics is characterized by permanent feed back loops between a social and a cognitive level. To be sure, although Marx frequently stressed the importance of mathematically formulated theories he never could think of transforming the venerable concept of dialectic into a mathematical framework.

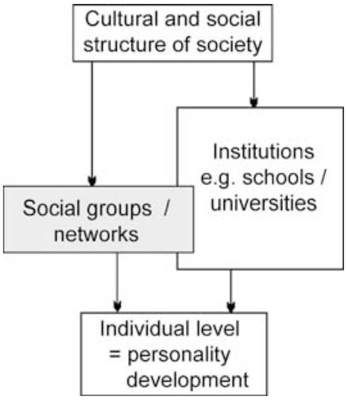
Marx had in principle captured the whole problem of social-cognitive complexity although only in an informal manner and formulated in classical philosophical terms. Yet his monumental approach was certainly not suited as a common foundation for the development of the social and cognitive sciences. Nearly at the same time the French sociologist Emile Durkheim, one of the founding fathers of modern sociology, postulated that social phenomena must only be explained by other *social* phenomena, which is one of the most important theoretical premises of the contemporary social sciences. As a consequence social scientists only deal with *social* complexity and consciously often neglect the influence of cognitive factors on social processes. Accordingly cognitive scientists only deal with *cognitive* complexity and neglect the impact of social factors on cognitive processes. Paradigmatic in this sense is the great work of Jean Piaget [34,35] who defined cognitive development as a principally autonomous process that can only be slowed down or accelerated by its social environment. With a famous expression of the German social theorist Niklas Luhmann one can say that the different disciplines of social-cognitive complexity reduced the complexity of their subject by concentrating on only one level of the two-level problem [26].

Matters became even more complicated when social theorists discovered that society is not a homogenous system but must itself be differentiated into different levels. The most basic differentiation is of course that of a micro-level and a macro-level [1,22]. In addition it is possible and sometimes necessary to introduce a meso-level, that is a middle range level of social reality. For example, if the micro-level is constituted by individual social actors or their actions respectively, if the macro-level is defined as a whole society or even an aggregation of societies like, e. g. the European Union, then a meso-level may consist of certain institutions, political parties and the like.

An additional complication and hence another increase of complexity must be accepted by taking account of the fundamental difference between social structure and culture. It is not possible for this article to mention all the numerous definitions of culture social theorists have developed in the last century. Following Habermas [15] and Giddens [12] it is sufficient to define the social structure of a society as the set of all social rules that are valid in a certain society and culture as the sum of all accepted knowledge, beliefs, world views, etc. in that society. By considering that fundamental distinction and by taking into account the different levels of society we apparently obtain a multi-dimension system, namely a system consisting of different levels that contain two dimensions – the socio-structural one and the cultural one. It is no wonder that social theorists concentrated on the social level of social-cognitive complexity, because this level is complex enough.

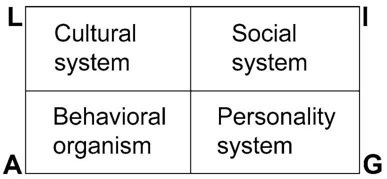
The main problem with such multi-level approaches is of course that the relations between the different levels and hence their interactions could usually be described only in an informal and frequently metaphorical manner. When a social theorist draws a graphical schema with different levels and additionally two dimensions the schema itself explains nothing but just states a problem. Somehow the levels interact and somehow individual actions are determined by the levels above, but in general this *somehow* cannot be precisely explained. To be sure, there are a lot of empirical studies that try, e. g., to analyze the impact of certain social structure on specific actions. But there is no general precise theory of links between the different levels of society [1].

Yet despite these difficulties in several fields of the social and cognitive sciences not only the different levels of society but also the cognitive level of individual actors is taken into account. That is especially the case in socio-psychology and in particular in theories of socialization. The analysis of processes of socialization indeed demands that cognitive levels and social levels alike must be considered, if one understands socialization as the process by which the individual mind is formed according to the social structure and the cultural norms and values of a particular society. Although in the general models of processes of socialization the relation between the social levels and the cognitive one is expressed only one-sided, i. e., only the determination of the personality by social structure is taken into account in contrast to Marx, at least it is suited to speak of attempts to capture the whole social-cognitive complexity of social reality. Such a model of socialization is shown in Fig. 1; it is a shortened version of a more complex model of Geulen and Hurrelmann [11].


Cultural and social structure of society
Institutions e.g. schools / universities
Social groups / networks
Individual level = personality development

Social Cognitive Complexity, Figure 1
A multi-level model of socialization

The well-known classical study of Berger and Luckmann "The Social Construction of Reality" is a paradigmatic example of such attempts [6] that in contrast to most one-sided models assumes a mutual interdependency of social and cognitive levels: According to them there is, in the tradition of Marx, a dialectical relation between the subjective social reality and the objective one. The subjective social reality consists of the internalized social rules and cultural values that constitute the personal identity, i. e. the individual personality. The objective social reality is generated by the according social actions and interactions of the socialized individual actors. Because the structure of this reality mirrors that of the individual personality the individual actions reproduce the objective reality, which was the determining factor of the constitution of the subjective reality. As in the case of Marx we have again a permanent feedback between the social and the cognitive level. Younger individuals participate in this process by taking over the structure of objective reality from their elders, by reproducing this structure in their own actions, and so forth. Although the theory of Berger and Luckmann does not deal with social progress, i. e., changes of social structure and culture, their model demonstrates how in principle it is possible to describe the interdependency



Social Cognitive Complexity, Figure 2
The general action system

of social and cognitive levels, even if only in an informal manner.

The most famous attempt to capture social-cognitive complexity in a whole model is certainly the well-known model of Parsons [8,29,32] of the differentiation of the action system. A simplified version is shown in Fig. 2.

The general assumption of Parsons is that each system is differentiated into four subsystems according to the famous AGIL-schema; the same logic of differentiation is valid for the subsystems and so forth. A means Adaptation, G stands for Goal Orientation, I is Integration, and L means Latent Pattern Maintenance. (The explanation of the AGIL-schema cannot be done in this article [29,33]). The picture shows that the social subsystems (social and cultural systems) constitute together with the cognitive subsystems (personality and behavioral system) the whole action system. The specific form of interaction between the four subsystems is called interpenetration, which roughly means that the four systems influence each other. For example, personalities are formed via socialization by the social and cultural subsystems, social and cultural systems are reproduced by personalities and according behavior and so on. Yet again we only have an informal theory, although the general logic of the AGIL-differentiation is certainly a great achievement.

To put it into a nutshell, the problem of dealing with social-cognitive complexity and not reducing the problem to one level is seen by many theorists. Yet the theoretical models are always just informal descriptions of the problem and programmatic foundations, if one looks at them from a precise, i. e. mathematical point of view.

Social-Cognitive Complexity, Algorithms, and Bottom Up Models

Despite many attempts to formulate the complexity of social-cognitive systems in a mathematical way for a long time it was not possible to transfer the mathematical tools of, e. g., mathematical physics to the problems of social reality. The use of statistical methods in empirical social research and experimental psychology is of course a well tested method since many decades. The problem was and frequently still is the mathematical formulation of *theoretical* models. Apparently the use of in particular differential equations is not suited for the problems one has to deal with in the social and cognitive realms. To be sure, simple cases can be analyzed in a mathematical way e .g. [5], but the complexity of social-cognitive systems in the sense of the great theorists mentioned in the preceding section could not be captured by models founded on the classical methods of the calculus.

The reasons for this are the peculiarities of social-cognitive systems. They are constituted, as was mentioned in the definition, by several levels that permanently interact and thus must be characterized by a multi-dimensional dynamics. In addition, social-cognitive systems are able to change their specific rules and structure by adapting to varying environmental conditions. It is of course possible to undertake a time series analysis in order to find some regularity in the historical development of such systems, but a time series analysis just describes a system's behavior in a phenomenological manner and gives no explanation. To make matters worse, even if one has found some regularity by a careful time series analysis the system may change its behavior via the varying of its rules; accordingly the time series will change and the analysis becomes worthless for the next time span. It is no wonder that more than once social and cognitive scientists alike believed that it would never be possible to develop a mathematical science of social and cognitive systems.

Fortunately things have changed by the development of new mathematical tools for the analysis of complex systems and in particular by the emergence of complex computer programs. These new mathematical modeling techniques, as cellular automata, Boolean networks, evolutionary algorithms, and artificial neural nets are on a first sight rather different methods but on a second sight they have much in common:

a) All modeling techniques are heuristically oriented to some natural processes as the reproductive capacity of living systems in the case of cellular automata and Boolean nets, the logic of biological evolution in the case of evolutionary algorithms, and the cognitive processes of the brain in the case of artificial neural nets. Sometimes these new mathematical tools are put together under the a bit misleading name of Soft Computing [48] but it would be more appropriate to call them nature oriented algorithms.
b) All these algorithms are very well suited for the construction of so-called bottom up models [24]. This term is used in contrast with top down models and means that the model construction is done by starting at the level of the elements of the system that shall be mod-

eled. Accordingly the units of the models represent the elements of the respective systems. The dynamics of the empirical system is represented by certain rules of interaction and the thus caused behavior of the whole model. The model's behavior is generated as an emergent consequence of the local interactions of the units of the model. This procedure allows constructing the respective model by *immediately* transferring the empirical observations of the system's elements and their rules of local interaction into the model. It is not necessary to make an abstract representation of the whole system by an aggregation of the elements.

For example, the modeling of a social group via a cellular automaton [21] can be done the following quite natural way: The cells of the cellular automaton represent the different individual members of the group and their specific rules of interaction determined by the group's social hierarchy are represented by the rules of transitions of the cellular automaton. It is then possible, at least in some cases, to predict the group's behavior by according simulation runs of the cellular automaton, as we did in several successful cases. The work of many researchers demonstrated that in particular cellular automata are very well suited for the analysis of social groups [13,16,25,30,39,41,43].

c) The capability of social-cognitive systems to adapt, i. e. to change their specific rules of interaction in order to fulfill certain environmental conditions can be modeled by the hybridization of the respective models, i. e., by the coupling of different modeling tools and such the constructing of a hybrid model [14]. In order to manage a successful adaptation adaptive systems must have not only rules of interaction but also meta rules, namely additional rules that operate on (*meta*) the rules of local interaction and change them according to the environmental demands. The best-known example for such adaptive capabilities is of course the changing of genomes by the genetic operators of mutation and recombination and by the force of natural selection in biological evolution. In this example the rules are those of individual epigenesis and ontogenesis; the meta rules are the genetic operators, steered by selection, that change the genome by variation.

A mathematical model of these processes of adaptation can be quite easily done by the construction of a hybrid system, consisting of several Boolean or logical nets respectively, representing the individual genotypes [18], and a genetic algorithm that operates on the rules of the Boolean nets and their specific topology.

Another example for adaptation is the case of political reforms. In nearly all advanced societies the *official* social rules of interaction are defined by laws and bureaucratic prescriptions. In parliamentary societies the meta rules are defined by the competence of the parliaments to change the laws and prescriptions by discussions and votes of majority. The meta rules then are given by the rules of procedure by which the parliaments have to operate. A variation of the rules of interaction by parliamentary decisions is done when the internal or external conditions of the society demand a changing of social structure. It can be demonstrated that democratic societies are more adaptive than dictatorships because dictatorial societies are both unwilling and unable to change their structure because of the resulting unrest and the danger for the ruling elite to lose its power [19].

An according model for such phases of political changing could, e. g. be constructed by the coupling of a cellular automaton (the model for the population) and an evolutionary strategy (the meta rules).

d) In contrast with a pure time series analysis bottom up models are able to explain, i. e. they can reduce observed phenomena to causes, namely to previous events that are causally linked to the observed event. In cases of complex systems such previous events can be either a previous state of the system. In that case the explanation consists of naming this particular previous state and the rules that generated the present state from the previous one. Or the previous event is an external disturbance of the system. Then the explanation consists of naming the disturbance and the reaction of the system's rules to the disturbance, i. e. to the externally changed state of the system. The third and most difficult case is that of an operation of the system's meta rules, their variation of the rules of interaction, and the generation of the present state via the changed rules of interaction. To be sure, the analysis and explanation of all different cases by the construction of an according valid model needs a lot of information about the empirical system. Yet a thorough explanation of such complex systems as socio-cognitive ones can only be done by the construction of models and simulation runs that use these new mathematical tools. To be sure, explanations of this kind can principally also be done by top down models and have been successfully done in physics and chemistry. Only the third case needs the distinction between rules and meta rules and hence demands models constructed the way sketched above.

Theoretical Foundations

The social sciences are, unfortunately, characterized by a multitude of different theoretical approaches that cannot be enumerated in this article. Yet for attempts to construct mathematical models of social or even social-cognitive processes there are two main approaches, namely the Rational Choice approach (abbreviated RC) e .g. [17] and a social systems theory approach, founded on social rules and structure. Both are based on certain anthropological foundations about the nature of man or social actors respectively.

The RC-approach is based on the assumption that man is basically an egoistical being who always tries to maximize his own profit without caring for others. This assumption has a venerable tradition and it goes at least back to the Leviathan of Hobbes, namely the question how social order is possible in a world of egoists. The term "rational" means that a social actor tries to find the optimal strategies in each action situation, i. e., he analyzes, which action as means could serve him best for his maximum welfare as goal. Hence each social action situation is considered by an actor as a situation in a strategic game like chess or poker: he rationally analyzes the different possible strategies and chooses the best – therefore rational choice.

The RC-approach has a great advantage for the purpose of mathematical model construction in the social sciences, namely the famous Theory of Games by von Neumann and Morgenstern [45]. Game theory was originally developed for the mathematical analysis of economical decision problems and already Max Weber stated that economical actions are a paradigm of rational actions, oriented to the most favorable proportion between different means and the goal of profit maximizing e .g. [46] (Oscar Morgenstern was an economist). But soon the partisans of RC-approaches generalized this idea and tried to capture the logic of social action in general by game theoretical models.

Usually game theoretical models are based on a so-called pay-off matrix. Because most RC-theorists distinguish only between two forms of behavior, namely cooperation and defection, a pay-off matrix is a two-dimensional one. A characteristic a pay-off matrix that is frequently used for the analysis of social dilemmas, in particular the famous Prisoner's Dilemma (PD), is the following one:

	C	D
C	3	0
D	5	1

C of course stands for cooperation and D for defection. As usual the matrix-values show the result of the combination of the two different action strategies. For example, the combination DC = 5 means that a defective player D gets five points while the cooperative player gets zero points (CD = 0). Accordingly a cooperative player gets three points if the other is also cooperative; the other gets three points too (CC = 3). The absolute values of course are arbitrary; the only condition for strategic games of the PD-type is that DC > CC > DD > CD.

In a famous study about *The Evolution of Cooperation* based on a tournament of different computer programs, i. e. programs that simulated winning strategies, Axelrod [2] showed that the best strategy was *Tit for Tat* (TFT). TFT can roughly be translated as "I do the same to you as you to me". TFT is a very simple strategy because the first player *A* always starts with a cooperative action and waits for the response of the other player *B*. If *B* is also cooperative then *A* keeps his cooperative actions. If *B* is defective then *A* switches his strategy and becomes defective too as long as *B* acts in a defective manner. The first cooperative action of *B* then is answered by *A* with an also cooperative action and so forth. Axelrod used for his own analysis a cellular automaton and confirmed his results later by using a genetic algorithm [3].

TFT is basically a cooperative strategy because each player who acts according to TFT always takes account of the strategies of his opponent and becomes defective only if his opponent does the same. Hence Axelrod concluded that the emergence of cooperation in a world of rational egoists is not only possible but also even probable because it can be mathematically shown that a cooperative strategy like TFT is the most favorable one for an egoistic rational player. In this sense the old question of Hobbes could finally be answered by the methods of mathematical game theory.

Despite the fact that the results of Axelrod are not so generally valid as Axelrod believed [9,19,31] it is no wonder that many partisans of the RC-approach saw these results as confirmation for the RC-approach and for the necessity to study social dilemmas by game theoretical methods. Sometimes it seemed that the social sciences had to be reduced to game theoretical analysis of decision problems. Yet these studies neglect several important difficulties.

Some time ago Lord Dahrendorf [7] formulated the important distinction between *homo oeconomicus* and *homo sociologicus*. *Homo oeconomicus* is the rational and egoistical actor, i. e. the basis of RC. *Homo sociologicus* in contrast is the social actor who orientates his actions to social rules and norms. He may be or not be an egoistical being but the important point is that his actions must be understood as the consequences of certain social rules that are valid in a specific situation. *Homo oeconomicus* understands a social action situation as a decision problem for finding the optimal strategy; his rationality, therefore, is the calculation for the optimal proportion between means and ends. The rationality of *homo sociologicus* on the other hand is an interpretative one: he must understand the situation in order to know, which rules determine the situation and which according actions he has to perform.

It is important to see that the question of Hobbes, which Axelrod and other partisans of RC tried to answer with a game theoretical analysis of social dilemmas, does not arise with the assumption of *homo sociologicus* as the basic nature of *homo sapiens*. Man is, according to this position, a social being by nature, i. e., his human nature is defined not only by characteristics like language or consciousness but also by his sociality. The basic question is then not how social order is possible at all – the question Hobbes answered by introducing the law enforcing state – but *how the construction* of social order is performed and how and why certain forms of social order were changed and others were not. Dahrendorf of course named this pair of concepts in this manner because he was convinced that *homo sociologicus* is the normal aspect of man as a social being; *homo oeconomicus* is a rather special case in situations where either no valid social rules exist, or where some actors deliberately decide *not to* follow the respective rules, or where social rules leave some space of freedom, which must be filled out by individual decisions.

It is not an arbitrary decision if one chooses the option for one anthropological option or the other. There are many reasons for the validity of the *homo sociologicus* position:

a) Studies in behavioral biology teach us that nearly all biological species consist of social beings, whether predators, prey, or plants; there are exceptions as, e. g., otters that live as loners but these are indeed exceptions. From anthropological studies it is to be learned that all our hominid ancestors lived in social communities and that socio-cultural evolution would not have been possible without human sociality [37]. Therefore, one may safely assume that human sociality is part of our biological heritage and not something humans had to create [47].
b) Social rules, and in particular social institutions, are important tools to reduce the complexity of the world and in particular that of the social world. Without accepted rules each new situation would be a new decision problem that has to be calculated anew. In contrast to that, social institutions offer a relief of the burden of new decisions because they enable actions by just following the rules [6,10,44]. That is particularly favorable if the rules have been proven in past situations. Hence, the evolution of social rules is more probable than Hobbes' famous *bellum omnium contra omnes* ("The war of all against all").
c) By following social rules it is possible to gain a social reputation as being reliable. Such a reputation frequently is favorable if others are undecided if they should trust an actor or not.
d) Finally, countless social studies have demonstrated that social actors mostly are rule-obeying beings. In a strict sense the social sciences could not have been as successful as they are if they would not have made the basic assumption of *homo sociologicus*.

These considerations do not mean that RC is wrong but that it is incomplete. There are certainly cases where social actors indeed act as egoistical rational players in a strategic game as for example in cases of social deviance or political revolutions. In particular, it is obviously a characteristic of modern democratic societies that the respective social rules nearly always leave a space of freedom for individual decisions. The social role of a teacher, for example, defines the necessity to act in a pedagogical way and to evaluate the efforts of the pupils. But it is an individual decision of the teacher if he mainly fosters the pupils or mainly acts as a critical evaluator of the pupils' performances. In this sense RC is an apt theory for some peculiarities of modern societies but it is not a general theory of human social action.

The consequence of these meta theoretical considerations is the following: The construction of mathematical models of social systems and of course of cognitive systems too must be founded on the basic logical unit of *rule*, either social rules or cognitive ones. To be sure, the elements of social systems are individual actors if one follows the bottom up modeling approach. Yet these actors are determined by the respective rules of their social systems; their actions, hence, must be understood as the result of the application of the rules valid in the respective situation. Certainly actors can be in error about the specific rules and/or they can decide to be socially deviant. In these cases it is particularly important to analyze the cognitive processes that lead the actor to social deviance. Yet the center of model construction and model interpretation must be the according rules.

We shall see in one example below that also in cases of social conformity the respective cognitive processes must be taken into account if one wants to understand how certain conformity emerged. Therefore, the interdependency of social and cognitive processes must be always be *principally* taken into account, although frequently it is possible to understand and even predict social behavior without considering the according cognitive processes.

A Universal Modeling Schema

The considerations of the preceding sections can be summarized in form of a universal modeling schema that captures the interdependency of different levels, the bottom up principle, and the logical dominance of the rule concept.

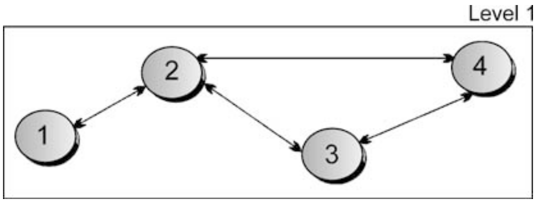
The basic *social* level of the schema consists of individual social actors who interact according to certain social rules. One may visualize this level, e. g., as the grid of a two-dimensional cellular automaton or the structure of a Boolean net. The level constitutes a dynamical complex system whose behavior depends, as usual, on the specific interaction rules, the initial states, and eventually operations of certain meta rules. Figure 3 shows such a level, which we may call level 1:

The concentration on this level and the neglect of a cognitive one apparently follows the mentioned methodical demand of Durkheim – social phenomena must only be explained by other social phenomena. Social actions are exclusively understood by the reconstruction of the respective rules that determined the actions. It is well known that many social processes can be understood and explained by taking into account only this level.

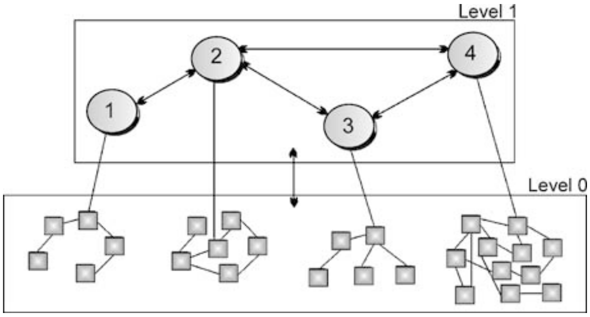
However, particularly in cases of social deviance, but not only there, the knowledge about the respective rules is not sufficient. Because social actors act with respect to certain rules *and* their particular thinking, worldviews etc. in many cases an according cognitive level must considered too. For example, the deeds of religious fanatics can only be understood if one reconstructs their fundamentalist worldviews and the according individual legitimization of criminal acts. In such cases social actors must also be considered as complex dynamical systems, i. e. cognitive ones with according cognitive rules and elements. The schema has to be extended as is shown in Fig. 4; the cognitive level may be called level 0.

Level 0 and level 1 already allow the modeling of many social processes, in particular those where an interplay of social and cognitive dynamics must be taken into account. Yet often even these two levels are not enough, as was mentioned in the introduction, because it is frequently necessary to model aggregations of individual actors. If, for example, one wants to analyze the political situation in a certain country with respect to the different political parties, then the concept of party of course means a certain institution consisting of specific rules and all party members. The interaction between such parties – coalitions, conflicts and so on – should be modeled on another level, namely level 2, the level of aggregations of actors (Fig. 5).

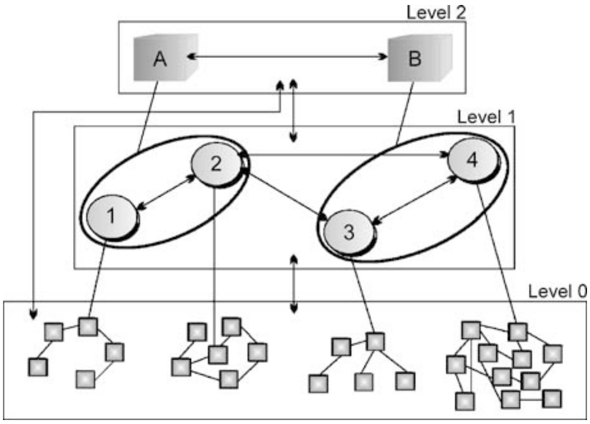
Note that level 2 must also be understood as a complex dynamical system with its own specific rules and elements. The elements there are usually called collective actors. All the three levels, hence, constitute certain complex systems whose dynamics is determined a) by the respective rules, elements, and initial states, b) by the eventual operation of meta rules that change the rules and in consequence the generation of states, and c) the interdependency between the different levels. This interdependency is visualized in



Social Cognitive Complexity, Figure 3
The social level as a complex dynamical system



Social Cognitive Complexity, Figure 4
Social actors modeled as cognitive complex systems



Social Cognitive Complexity, Figure 5
Social actors (level 1) modeled as cognitive complex systems (level 0) and as aggregations (level 2)

the pictures by the arrows that represent the mutual influences.

In contrast to many schemas of multi level models, e. g. the schema shown in the introduction, it must be noted that there is *in principle* a permanent feedback between all different levels. The action of individual actors (level 1) generate and reproduce the collective actors on level 2, yet collective actors in turn influence and determine the individual actors. A member of a political party acts *as* a party member. Accordingly the cognitive processes of individual actors determine their actions but these cognitive processes in turn are influenced by the result of the actions on the one hand and the collective actors the individual actor is a part of on the other. A party member not only acts but also thinks in concepts of the respective party. In addition, individual cognitive processes may directly influence collective actors, particularly in times of social and/or political change. History is full of the impact of individual cognitive processes on whole institutions, if the individual actors operate as reformers or revolutionaries.

It is certainly possible to extend even the schema shown in Fig. 5 by either adding a third social level, consisting of aggregations of collective actors, or by adding a second cognitive level by distinguishing between, e. g. conscious and unconscious thinking. Such extensions would lead to four- or even five-level models. Yet in most cases of social-cognitive modeling it is sufficient to operate with one, two or three levels. To be sure, it is always a question of the specific research interest how many levels and which ones one must take account of. The schema, after all, is just a *schema*, that is a frame work to guide the modeling processes when dealing with socio-cognitive processes – and not only these. The construction of a certain concrete model for a specific problem might be still a difficult task, in particular if one has to evaluate the validity of the specific model.

It is rather evident that this modeling schema can practically be applied to all complex social-cognitive processes. Yet the schema is in an even more ambitious sense a *universal* schema. This is a consequence from the mathematical characteristics of the algorithms mentioned in Sect. "Social-Cognitive Complexity, Algorithms, and Bottom Up Models". Imagine for example that level 1 is modeled by a cellular automaton, level 0 by one or several types of artificial neural networks, and level 3 again by a cellular automaton or a Boolean net. It is known from mathematical investigations that cellular automata are potentially equivalent to Universal Turing machines [36], which roughly means that no formal system is more general than such systems. This is the so-called Church–Turing thesis, which no mathematician seriously doubts. A corollary, the *physical* Church–Turing thesis, states that each physical, i. e. empirical system can be modeled via the use of a suited Universal Turing machine or an equivalent system. Although the physical Church–Turing thesis has not been proven in an exact mathematical manner nobody doubts this thesis either.

Without great difficulties it can be shown that each neural net and each Boolean net is equivalent to an according cellular automaton – the mappings between cellular automaton and the two classes of nets are injective. Hence principally each complex system can be modeled by using this schema *and* by using one or several of the mentioned algorithms. An important consequence from these considerations is the fact that a mathematical science of social-cognitive systems is not only possible but also that there is for each problem a general way how to treat it in a mathematical manner. The schema, therefore, can be understood as a constructive mapping from socio-cognitive phenomena to mathematical structures. The universality of the modeling schema is a constructive proof of the possibility of mathematical social and cognitive sciences.

Examples

These considerations show that and how mathematical models of social-cognitive processes can be constructed. Yet the most convincing proof is of course the demonstration of concrete examples, i. e., the models of specific complex social-cognitive systems. More detailed descriptions of these and other examples can be found in [20] and [21].

The Differentiating of Social Groups into Subgroups

A certain social group that consist of more than, say, three or four members is usually not a homogenous entity but is differentiated into several subgroups. The criteria for the generation of such subgroups are, of course, not always the same. In working teams for example the different qualifications of the members might be a criterion for the forming of specialized sub-teams. Yet in many informal groups a principle is certainly valid that may be called Homans' Principle [17] and that can be stated as follows: "Members of a social group favor interactions with people they like and they avoid interactions, if possible, with people they do not like." Everyday experience shows that this principle is indeed in many social groups an important criterion.

Since the foundation of sociometry by Moreno [28] the usage of a so-called sociomatrix or Moreno matrix respectively for the representation of social relations has become customary. If one wants to represent the mutual feelings of the members of a certain group via a sociomatrix the matrix may be:

	a	*b*	*c*
a	0	1	−1
b	1	0	1
c	−1	0	0

The matrix is coded with three values. Value 1 means a positive feeling or a liking respectively, value 0 means neutral or indifferent feelings, and −1 means negative feelings or disliking respectively. Hence the matrix can be understood as member *a* likes *b* and dislikes *c*, member *b* likes *a* and *c*, and member *c* dislikes *a* and is indifferent to *b*. The usage of such a sociomatrix as a mirror of the emotional structure of a certain group can be very important for, e. g. teachers, group leaders, group therapists or trainers of sport teams.

According to the Homans principle our research group COBASC constructed a cellular automaton with the goal to simulate and predict the emerging of different subgroups in specific groups, which we called in honor to Moreno the Moreno-CA (COBASC is the abbreviation for *Computer Based Analysis of Social Complexity*). It is a two-dimensional cellular automaton whose cells represent the members of the particular group that shall be modeled. The state of the cells represent the emotional states of the group member that range from "very dissatisfied" to "very satisfied." The cells state is dependent of the eight adjacent cells (a "Moore neighborhood") in the following way: the values of the emotional relations of the center cell to the eight cells of the Moore neighborhood are summed up; the emotional state of the center cell then is the arithmetical mean of these values. (The cells are usually modeled as squares; hence there are eight adjacent cells). The values of the emotional relations are coded as 1, 0, or −1 as in the example above. The task of each cell is to find a Moore neighborhood where its own emotional state is a maximum. Accordingly the rules of the Moreno-CA are:

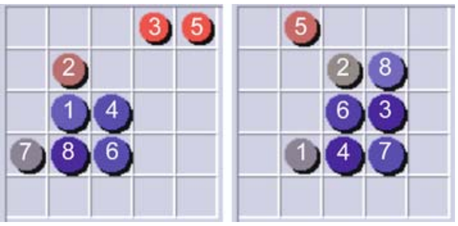
a) Place as many cells on the grid as there are group members; the rest of the grid's cells are free places.
b) Compute the emotional state for each "member" cell; vacant cells get the value 0.
c) Let each cell look on the grid for a Moore neighborhood where it can obtain a maximum state value.
d) Move each cell into that neighborhood.
e) If an attractor state, preferably a point attractor, has been reached, stop; if no attractor state could be reached after a certain number of runs, stop too. (A point attractor is a state the system will not leave although the rules are still operating; an attractor of longer period consists of several different states that the system will generate in cyclic order.)

The last rule is necessary because it can happen that the group reaches no attractor state at all or only those with rather long periods [21]. Factually the rules are a bit more complicated but this simplified version is the essence.

We validated this model with different social groups and obtained in all cases very satisfactory results. One empirical example shall serve to illustrate the methodical procedure:

With the permission of the teacher we asked a group of eight pupils of a primary school, with which pupils they would like to share a room in a youth hostel, with which pupils they would not share a room under any circumstances, and with which pupils they would share or not, according to the circumstances in the youth hostel. (It is often problematic to ask children (in this case 10 years old) directly if they like or dislike other children. That is why we chose this procedure of asking). These answers were transferred in the respective sociomatrix of the group. Afterwards the pupils were ordered to go into a strange classroom and choose places in order to wait for the teacher. The hypothesis was of course that the pupils would choose places according to their emotional relations to the other pupils. Then the values of the factual sociomatrix were inserted into the Moreno-CA; it reached a point attractor after 8 runs. The prediction of the program was then compared with the factual sitting order of the pupils (Fig. 6).

The program of course did not predict the absolute spatial distribution of the pupils in the classroom, for example if one pupil prefers to sit at a window. The Moreno-CA just predicts the relative positions of the pupils, i. e., which pupils are sitting together or which pupils are outsiders and are sitting alone. A comparison of the factual distribution and the prediction of the cellular automaton shows that the program indeed correctly prognosticated the sitting order for nearly all pupils with the exception of



Social Cognitive Complexity, Figure 6
The factual sitting order of the pupils (*left side*) and the prediction of the Moreno-CA

pupil 3. He is factually sitting besides the outsider pupil 5. However, the teacher of the children could give an explanation for this error of the program: Pupil 3 is known to be a practical jester who wanted to win the applause of the other children by teasing the outsider. Such melancholy character traits of course the program could not know.

We inserted the same sociomatrix into another program, namely a so-called self-organizing map or Kohonen feature map respectively (see next example). This is a special type of artificial neural nets. This net obtained even better results than the Moreno-CA because it placed pupil 3 beside pupil 5. An explanation for this result is given in [21].

To be sure, we did not want to predict just sitting orders of children in a classroom. The social experiment and the simulation by the two programs should show which subgroups existed in the whole group of children; this differentiation would become visible in the little experiment. The teacher, by the way, confirmed the results of the experiment and the prediction because according to her thorough knowledge the group of pupils factually consisted of an outsider and a homogenous group of all other pupils.

With respect to the modeling schema of the previous section this model uses only level 1, namely the social one. To be sure, in an implicit way, the emotional state of the children also plays a decisive role: there are no explicit social rules in this situation and the desire of the children to be in the neighborhood of their friends and to avoid other children is, according to Homans' principle, the main determination of the factual behavior. But in the model it was sufficient to represent the children as finite state automata, i. e., units whose state values are a direct consequence of their respective social milieu – in the model the Moore neighborhood. The Moreno-CA and the social experiment with the children demonstrate that indeed in many cases it is sufficient to construct models with only one level, in particular the social one.

Socialization by Learning from a Model

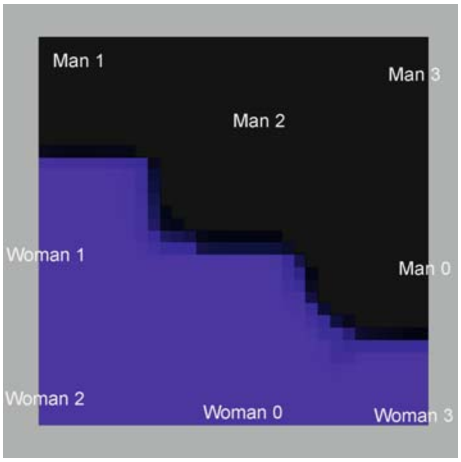
In Sect. "Introduction" it was demonstrated why theories of socialization always must take account of both social level(s) and the cognitive one. The social level represents the social milieu that is decisive for the specific socialization process; if one wants to explain the particular characteristics of this milieu one must consider additional social levels. The cognitive level of course represents the personality or identity of the individual that is formed or influenced by the respective milieu. The term formed though does not necessarily mean that socialization is a passive process. The socialized individual has the task to actively construct a worldview consisting of cultural values, social rules, and a culture specific knowledge.

One of the most known theories of socialization is the theory of learning from a model by Albert Bandura [4]. This theory states that socialization is to a high degree dependent on other people in the social milieu, who can act as models, i. e. as paradigms. The socialized individual perceives these models as positive or negative examples for its form of thinking and behavior. In other words, young people and children do not learn cultural values and specific forms of social behavior by internalizing general rules but by concrete examples they can observe and imitate. (There is a striking similarity between the socialization theory of Bandura and the cognitive theory of prototypes by Eleanor Rosch [23], which can only be mentioned here.)

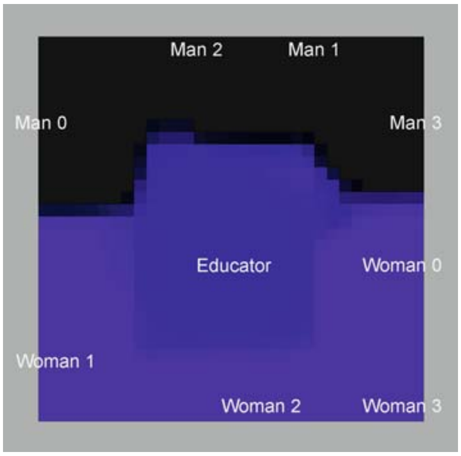
We applied this theory to a real case, namely a male youth from the *Ruhrgebiet*, an industrial area in the West of Germany. The youth, who shall be named Tom, of course a fictitious name, grew up in a workers family; he told us his story in an interview that was performed by one of our students. The father was an aggressive and violent man who often hit Tom's mother and frequently got involved in brawls and fights in the pubs. The mother was a submissive person who tried to keep up the family by working as a cleaner. Tom, therefore, had to care for his two younger sisters, which he very much disliked although he liked his sisters. When Tom was 10 years old his father left the family and disappeared. At the time of the interview Tom was 17 years old and had never heard any more from his father. Because Tom joined a gang of hooligans and got into trouble with the authorities and because the mother was unable to care for the two girls and Tom alike, Tom was admitted to a hostel of delinquent youths.

Despite the behavior of the father Tom adored him because *the father was a real man*. "Real" men are not afraid of other men, they become violent and hit if they feel insulted and in particular they dominate women, if necessary also by violence. "Real" women, on the other hand, are submissive and obedient to their men; they cook, wash and care for the children while real men work outside the house. As a consequence Tom developed a dichotomous worldview: men are strong and women are weak; women should obey and men command. The worlds of men and women must not be mixed in the sense that women take over tasks of men and men in turn care for the household. Consequently Tom despises the educators in the hostel because the female educators try to give orders to young men like Tom and the male educators perform household tasks like cleaning and cooking.

Tom apparently obtained his worldview by learning from the positive model of his father – "I want to be a man



Social Cognitive Complexity, Figure 7
The reconstruction of a dichotomous worldview: female versus male clusters



Social Cognitive Complexity, Figure 8
Increasing the female cluster by adding other "real women" and "false" men

like my father" – and the negative model of his mother. He also remarked that most men he met in his childhood were similar to his father and the women were similar to his mother. Hence he was early influenced by a homogenous milieu, where men and women all occupied the same gender specific roles.

For a simulation of Tom's biography with the result of a dichotomous worldview we used a Self-Organizing Map (SOM) or Kohonen Feature Map respectively that belongs to the class of neural networks, which learn in a non-supervised manner. In contrast to the forms of supervised learning non-supervised learning is not determined by externally given targets but the learning process is characterized by an internal logic. To put it into a nutshell, non-supervised learning is the ordering of a set of data according to certain internal criteria. There are different types of SOMs; we used a so-called Ritter–Kohonen model [38]. The technical details do not matter here; they can be looked up in any textbook on neural nets. The main intention of the simulation was to reconstruct the genesis of Tom's worldview and to demonstrate the according behavior in the present. Hence we inserted formal representations of the father and the mother as positive and negative models and inserted additional persons into the SOM. The SOM then constructed clusters of the different persons who are distinguished according to the criterion "real men" versus "real women". A first result is shown in Fig. 7.

The spatial nearness on the screen represents similarities or differences between the clustered persons respectively.

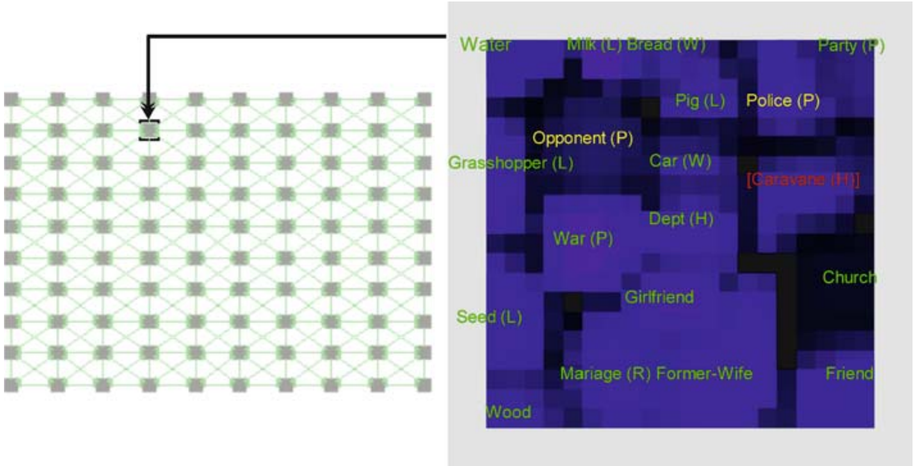
Afterwards we inserted other persons who differed from the first in certain personal respects. In particular we inserted formal representations of male educators who had female characteristics like cooking or cleaning. The SOM placed these men near the female cluster according to Tom's worldview: Tom does not accept these educators as real men but sees them more as women in male disguise. This is shown in Fig. 8.

The SOM apparently is able to simulate the important parts of Tom's socialization processes. As a SOM is a purely deterministic algorithm the successful simulation of Tom's cognitive genesis must be seen as a deterministic process too. Tom, so to speak, had no chance to become another person because the social circumstances determined his biography in a complete way.

The model uses only level 0, i. e., the cognitive one. Yet the social level is implicitly present too because of the inputs we inserted to the program; these inputs represent the social milieus in which Tom grew up and presently lives. In addition, the worldview of Tom has certainly social consequences not only for his biography. His membership in a hooligan gang, where he feels at home, indicates that his biography is representative for many youth members of such gangs. The according consequences for societal problems are obvious. In this respect the example of Tom also contains level 2, e. g. the level of collective actors like the different hooligan gangs and the social institutions that have to deal with them.

Modeling Social-Cultural Evolution

The last example is a rather complex model whose theoretical foundations and historical arguments can be read up in [20]. In contrast to the two preceding examples that be-



Social Cognitive Complexity, Figure 9
Cognitive state of an individual actor and his place in the artificial society

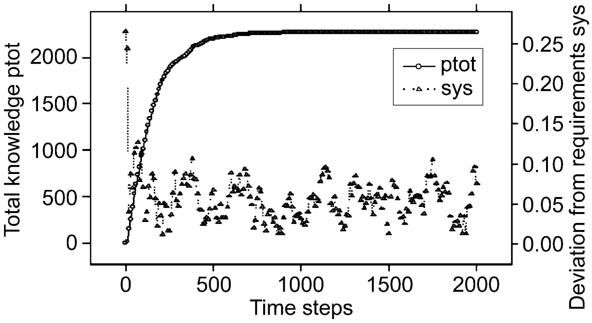
long to the field of micro-sociology this model deals with the evolution of whole societies. The basic assumption is that socio-cultural evolution is generated by creative inventions of individual actors and by the acceptance of these new ideas by the respective society. The social structure and the cultural norms of a society may promote or hinder the generation and acceptance of new ideas. In particular, the decisive factor is the degree of role autonomy for the occupants of creative roles like engineers, artisans, artists or scientists. If this degree is sufficiently high then the respective society evolves, i. e. increases its cultural capacity and generates social structures that are more efficient to deal with the problems of the society than old ones. If the degree is low, then the society stagnates, i. e. it becomes caught in a cultural and socio-structural attractor.

Comparative studies of the history of many societies always showed the same results, namely an early blossoming and cultural growth, then a period of consolidation, and finally stagnation or even regression, i. e. the return to some previous states [40,42]. According to our theoretical assumptions in these cases the degree of role autonomy was not high enough to allow an unhindered evolution. Only one case is known where apparently cultural evolution did not get caught in an attractor, namely the culture of European Modernity that was the historical basis for the contemporary Western culture. Historical studies confirm our hypothesis: In all cases of socio-cultural evolution the degree of role autonomy was much lower than in the European Modernity and the Western culture in general.

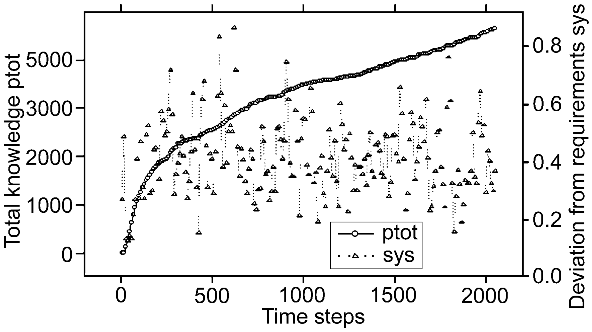
To validate these considerations by the use of a mathematical model we constructed the so-called socio-cultural-cognitive algorithm (SCCA). The model consists of a) a social level that is constructed as a cellular automaton. As usual the cells represent individual actors, i. e. members of a certain society. The rules of the cellular automaton determine the degree of role autonomy, i. e. the proportional ability of the actors to learn and create new ideas on the second level. b) The second level is a cognitive one: Each cell consists of a combination of different neural networks, namely several so-called bi-directional associative nets (BAM) and a SOM from the type used in example "Socialization by Learning from a Model". The task of the BAM is to associate observed characteristics from certain objects with the according semantic concepts – small, furry, and meowing is associated with cat. The task of the SOM is to order the concepts into a semantical network. These cognitive processes can be performed either in the way of learning, i. e. taking over new concepts from other actors, or by creating new concepts. The sum of the results of the individual cognitive processes is defined as the state or level that the respective culture has reached. Figure 9 shows an individual actor and its place on the grid of the cellular automaton.

The degree of role autonomy is inserted into the rules of the cellular automaton. We did many experiments with different values of this degree. The most important result is that in most cases the according culture indeed showed an evolutionary pattern known from human history, in particular described by the great British historian Arnold Toynbee (Fig. 10).

Only in the few cases with a high degree of role autonomy our model showed an evolutionary path known from the history of European Modernity. This is shown in Fig. 11.



Social Cognitive Complexity, Figure 10
A normal path (Toynbee path) of cultural evolution



Social Cognitive Complexity, Figure 11
A "Western development"

Societies with a sufficient high degree of role autonomy are not hindered in their cultural development although of course such societies may be stopped in their evolution by external forces. That could have been the reason for the final stagnation of the ancient Greek culture that also was characterized by a high degree of role autonomy for the occupants of the creative roles. Our model confirms that such societies are very seldom, which may be indeed the main reason for the fact, as Toynbee [42] observed, that the overwhelming majority of historically known societies stagnation was their inevitable fate.

The model is characterized by interplay of the two levels with consequences for a third one because the social structure determines the individual creative processes that in turn determine the fate of the whole society. In addition, sufficient developments on the cognitive level may also change the social structure: the better the cognitive improvements are, the higher becomes the degree of role autonomy on the social level. Such a rule is confirmed too by historical observations of the European history: In the Middle Ages the degree of role autonomy was not much higher than in contemporary cultures like feudal China or the Islamic societies. By the cultural progress during European Modernity the degree of role autonomy permanently increased. The model, hence, shows the consequences of such interdependent dynamics of the different levels that may explain important courses of history with respect to the fate of whole cultures.

Conclusions and Future Directions

The general possibility of a mathematically formulated science of social and cognitive processes was demonstrated in the Universal Modeling Schema and the fact that suited formal models like cellular automata or neural networks are equivalent to Universal Turing Machines. How many levels one needs for a concrete model, which ones, and how the relations between the levels are to be formulated is of course a question of the specific research problem. Constructing models of socio-cognitive processes will never be a trivial task. The modeling schema gives general orientations, but the detailed knowledge of the problem and in particular of the empirical data is always a necessary condition for the construction of social-cognitive models. That is of course a truism if it is stated in such a general manner but one that must always be taken into account. In particular, the knowledge about model construction and of suited algorithms never substitutes the profound knowledge about the specific fields of research.

The general relation between a social level and a cognitive one can also be stated in terms of individual actors. If one takes for example a most basic social situation, i. e. the communicative interaction between an actor A and another actor B, and if one assumes that during the interaction neither the social rules of this situation nor the cognitive processes of the actors change, then this interaction can be written the following way:

If f designates the set of social rules decisive for this situation and if g_A and g_B designate the cognitive processes of A and B respectively, then

$$f(A) = g_B(B)\,.$$

In other words, the application of the social rules f on the personality of A causes A to send an appropriate message to B, who in turn is moved to perform the cognitive processes g_B. In this case the social level determines the cognitive one, as is certainly the case in most forms of everyday communication. Accordingly the equations can be written if the cognitive processes determine the social rules, for example if a student becomes more intelligent than his professor by the teaching of the professor. (For general formulations of these equations see also [21]).

The most important problem for a socio-cognitive science that operates with mathematical models and according computer simulations is without doubt the proper relation between theory, formal model and simulation program. A complete research program that has to take care of all three components follows the order: theory construction or theory selection respectively – transformation of the theory into a suited model – construction of a simulation program – comparing the results of the simulation runs with known empirical data. If everything went right, the last step will be a confirmation of the whole process. Yet since Murphy's law we know that each step may contain errors and if the data are not compatible with the simulation results then the error(s) may be in the theory, the model, the simulation program, or in several of the steps.

In particular the fundamental role of theory has frequently been neglected in the many attempts to construct a mathematical or computational respectively social science by using computer models of socio-cognitive phenomena. The mentioned dominance of RC-models on the foundations of Game Theory is just one indicator for this problem. Yet the social and cognitive sciences are much more than theories about rational and egoistical players in strategic games. Therefore, the most important problem for the future is the careful analysis of the research program formulated in the three or four steps mentioned above. Only then the use of mathematical models and suited simulation programs will be the methodical basis for a real socio-cognitive science.

Yet we have no alternative to the development of such sciences. If this research program is not successful we will always be the unconscious slaves of the social conditions we produced ourselves, in particular of a future we did not wish for. A science of the complex socio-cognitive reality is certainly not a sufficient condition for the solving of all our social problems. But it is a necessary condition for the task that we must understand our own species, that is to understand our own social and cognitive processes.

Bibliography

Primary Literature

1. Alexander JC, Giesen B, Münch R, Smelser NJ (eds) (1987) The Micro-Macro-Link. University of California Press, Berkeley
2. Axelrod R (1984) The Evolution of Cooperation. Basic Books, New York
3. Axelrod R (1987) The Evolution of Strategies in the Iterated Prisoner's Dilemma. In: Davis L (ed) Genetic Algorithms and Simulated Annealing. Morgan Kauffman, Los Altos
4. Bandura A (1986) Social Foundations of Thought and Action. A Social Cognitive Theory. Prentice Hall, Englewoods Cliff
5. Beltrami E (1993) Mathematical models in the social and biological sciences. Jones and Bartlett, London
6. Berger P, Luckmann T (1966) The Social Construction of Reality. Doubleday, New York
7. Dahrendorf Lord R (1958) Homo Sociologicus. Ein Versuch zur Geschichte. Bedeutung und Kritik der Kategorie der sozialen Rolle. Leske Buderich, Opladen
8. Fararo TJ (2001) Social Action Systems. Foundation and Synthesis in Sociological Theory. Praeger, London
9. Fogel DB (1993) Evolving Behaviors in the Iterated Prizoner's Dilemma. Evol Comput 1:77–97
10. Gehlen A (1956) Urmensch und Spätkultur. Frobenius, Bonn
11. Geulen D, Hurrelmann K (1982) Zur Programmatik einer umfassenden Sozialisationstheorie. In: Hurrelmann K, Ulich K (eds) Handbuch der Sozialisationsforschung. Beltz, Weinheim, pp 5–67
12. Giddens A (1984) The Constitution of Society. Outline of the Theory of Structuration. Polity Press, Cambridge
13. Gilbert N, Troitzsch KG (1999) Simulation for the Social Scientist. Open University Press, Buckingham
14. Goonatilake S, Khebbal S (eds) (1995) Intelligent Hybrid Systems. Wiley, London
15. Habermas J (1981) Theorie des kommunikativen Handelns, vol 2 (Theory of communicative action). Suhrkamp, Frankfurt
16. Hegselmann R, Mueller U, Troitzsch KG (eds) (1996) Modeling and Simulation in the Social Sciences from the Philosophy of Science Point of View. Kluwer, Dordrecht
17. Homans GC (1950) The Human Group. Harcourt, Brace, Jovanovich, New York
18. Kauffman S (1993) The Origins of Order. Oxford University Press, London
19. Klüver J (2000) The Dynamics and Evolution of Social Systems. Kluwer, Dordrecht
20. Klüver J (2002) An Essay Concerning Socio-Cultural Evolution. Theoretical Principles and Mathematical Models. Kluwer, Dordrecht
21. Klüver J, Klüver C (2007) On communication. An Interdisciplinary and Mathematical Approach. Springer, Dordrecht
22. Knorr-Cetina K, Cicourel AV (eds) (1981) Advances in Social Theory and Methodology. Toward an Integration of Micro- and Macro-Sociologies. Routledge and Kegan Paul, Boston
23. Lakoff G (1987) Women, Fire, and Dangerous Things. What Categories reveal about the Mind. The University of Chicago Press, Chicago
24. Langton C (1988) Preface. In: Langton C (ed) Artificial Life. Addison-Wesley, Redwood
25. Levy S (1993) Artificial Life. Penguin Books, London
26. Luhmann N (1984) Soziale Systeme. Suhrkamp, Frankfurt
27. Marx K, Engels F (1969) Deutsche Ideologie (German Ideology). MEW 3
28. Moreno JL (1953) Who shall survive? Foundations of sociometry, group psychotherapy and sociodrama. Beacon House Inc., Beacon
29. Münch R (1988) Theorie des Handelns. Suhrkamp, Frankfurt
30. Nowak A, Levenstein M (1996) Modeling Social Change with Cellular Automata. In: Hegselmann R, Trotzsch K, Muller U (eds) Computer simulations from the philosophy of science point of view. Kluwer, Dordrecht
31. Nowak MA, May RM (1993) The Spatial Dilemma of Evolution. Int J Bifurc Chaos 3:35–78
32. Parsons T (1968) The Structure of Social Action. Academic Press, New York
33. Parsons T, Platt GM (1973) The American University. Harvard University Press, Cambridge

34. Piaget J (1972) The Principles of Genetic Epistemology. Routledge, London
35. Piaget J (1972) The psychology of intelligence. Littlefield Adams, Totowa
36. Rasmussen S, Knudsen C, Feldberg R (1992) Dynamics of Programmable Matter. In: Langton C, Taylor C, Farmer JD, Rasmussen S (eds) Artificial Life II. Addison-Wesley, Redwood
37. Read DW (2005) Change in the Form of Evolution: Transition from Primate to Hominid Form of Social Organization. J Math Sociol, Special Issue Emergence Soc Order 25:3
38. Ritter H, Kohonen T (1989) Self-organizing Semantic Maps. Biol Cybern 61:241–254
39. Schelling TC (1971) Dynamical Models of Segregations. J Math Sociol 1:143–186
40. Spengler O (1926) The Decline of the West. Deutsches Original: (1922) Der Untergang des Abendlandes. Alfred Knopf, New York
41. Suleiman R, Trotzsch KG, Gilbert N (eds) (2000) Tools and Techniques for Social Science Simulation. Physica, Heidelberg
42. Toynbee A (1934–1961) A Study of History (12 vols). Oxford University Press, Oxford
43. Troitzsch KG, Mueller U, Gilbert N, Doran J (eds) (1996) Social Science Microsimulation. Springer, Heidelberg
44. Vanberg VJ (1994) Rules and choice in economics. Routledge, New York
45. von Neumann J, Morgenstern O (1944) Theory of Games and Economic Behavior. Princeton University Press, Princeton
46. Weber M (1981) Die protestantische Ethik. Mohr, Tübingen
47. Weingart P, Mitchell SD, Richerson PJ, Maasen S (eds) (1997) Human by Nature. Between Biology and the Social Sciences. Lawrence Erlbaum, Mahwah
48. Zadeh LA (1994) Fuzzy Logic, Neural Networks and Soft Computing. Comun of the ACM 37:77–84

Books and Reviews

Alexander JC (1983) Theoretical Logic in Sociology, vol 3. The Classical Attempt at Theoretical Synthesis: Max Weber. Melbourne-Henly, London

Brennan G, Buchanan M (1985) The Reason of Rules. Constitutional Political Economy. Cambridge University Press, Cambridge

Coleman JS (1974) Power and the Structure of Society. Norton, New York

Coleman JS (1990) Foundations of Social Theory. Cambridge, London

Diekmann A, Schmidt P (1998) Environmental Sociology, Special Issue of Rationality and Society. Sage, London

Ginsberg A (1990) Connecting Diversification to Performance. A Sociocogn Approach. The Acad Manag Rev 15:3:514–535

Hamilton D, Devine P, Ostrom T (eds) (1994) Social Cognition: Impact on Social Psychology. Academic Press, London

Howard J (1994) A social cognitive conception of social structure. Soc Psychol Q 57:210–27

Howard RW (1995) Learning and Memory. Major Ideas, Principles, Issues and Applications. Praeger, Westport

Parsons T (1951) The Social System. The Free Press of Glencoe, Glencoe

Resnick LB, Levine JM, Reasley SD (ed) (1991) Perspectives on socially shared cognition. American Psychological Association, Washington

Rogoff B (1990) Apprenticeship in thinking: Cognitive development in social context. Oxford University Press, New York

Schech S, Haggis J (2000) Culture and Development. A Critical Introduction. Blackwell Publishers, Malden

Singley MK, Anderson JR (eds) (1989) The transfer of cognitive skill. Harvard University Press, Cambridge

Snijders T, Snijders TAB, Bosker R (2000) Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling. Sage Publications, London

Walsh DM (ed) (2001) Naturalism, Evolution and Mind. University Press, Cambridge

Wertsch JV (ed) (1985) Culture, communication and cognition: Vygotskian perspectives. Cambridge University Press, Cambridge

Winch C (1998) The Philosophy of Human Learning. Routledge, London

Social Coordination, from the Perspective of Coordination Dynamics

Olivier Oullier[1,2], James A. S. Kelso[2]

[1] Laboratoire de Neurobiologie Humaine (UMR 6149), Aix-Marseille Université, Marseille, France

[2] Human Brain and Behavior Laboratory, Center for Complex Systems and Brain Sciences, Florida Atlantic University, Boca Raton, USA

Article Outline

Glossary

Self-organization Self-organization lies behind all structure and pattern formation in nature's complex systems, including the human brain. Self-organization is a principle governing a system where no agent-like entity is ordering the elements, telling them where and what to do. In self-organizing systems, low-dimensional dynamics are revealed by changing one (or more) control parameter(s) whose role is simply to move the system through a series of state changes without prescribing its behavioral patterns.

Coordination dynamics Coordination dynamics seeks the laws, principles and mechanisms underlying the coordinated behavior of different kinds of components at multiple levels of description (molecules, cells, circuits, etc). It is an overarching conceptual framework that describes, explains and predicts how patterns of coordination form and change at multiple levels of brain and behavior. The brain, mind and behavior are linked by virtue of sharing a common underlying coordination dynamics.

Information exchange A remarkable fact is that in contrast to classical dynamics that deal with fundamental quantities such as mass, length and time and their relations, coordination dynamics is informational in nature, dealing with informational quantities of a relational kind that couple different parts of a system or different systems.

Phase transitions Phase transitions are the true illustration that a system is self-organizing. They are spontaneous qualitative pattern changes occurring as parameters are changed quantitatively. When they occur, abrupt switches from one coordinated pattern to another are observed and the dynamics of the entire self-organizing system is dominated by one or a few collective variables: the order parameters.

Stability Stability is a key concept in coordination dynamics. Here the stability is of coordination or collective variables. The (loss of) stability of a self-organizing system indicates whether a phase transition is to occur. In order to evaluate the stability of a system, one can perturb it and measure the time it takes for the system to return to its initial state, i. e. its *relaxation time*. A number of other converging measures have been used to measure stability in coordination dynamics such as *switching time* (the time it takes for the system to switch from one pattern to another when phase transitions occur) and *critical fluctuations* (the increase of variability of the collective variable in the vicinity of the phase transition).

Definition of the Subject

Social Coordination Dynamics (SCD) explores, at both behavioral and neural levels, the mechanisms mediating the formation and dissolution of bonds between individuals. SCD applies the concepts, methods and tools of informationally coupled self-organizing systems (coordination dynamics) to quantify real time social processes. Just as coordination dynamics deals with how the parts of complex systems work together in a meaningful way to achieve goals, so SCD aims to understand the interplay of forces operating at both individual and collective levels to produce effective social behavior. SCD offers a novel perspective and new metrics to explore systematically a fundamental form of human bonding (or lack thereof), and the self-organizing processes that underlie its persistence and change over space and time. SCD therefore complements recent developments in several fields such as sociology, social cognitive neuroscience, behavioral economics, game theory and neuroeconomics.

Introduction

Coordination can be broadly defined as a functional ordering among interacting components in space and time. Coming in many guises, coordination represents one of the most striking features of living organisms. The science of coordination, *Coordination Dynamics* (abbreviated CD) [36,37,38,41] stems from a complex systems framework based on the theory and methods of informationally coupled self-organizing dynamical systems (see ► Coordination Dynamics). CD explores a number of basic coordination phenomena that cut across a wide range of levels, creatures and functions. Of particular relevance to social coordination are: (i) patterned states of coordination remain stable in time despite perturbations; (ii) component parts and processes (dis)engage in a flexible fashion depending on functional demands and/or changes in environmental conditions; (iii) multiple coordination states exist rendering living things *multi-functional*, effectively satisfying the same (or different) set of circumstances; (iv) switching from partially to fully coordinated states and vice versa is commonplace; (v) selection of coordination patterns is tailored to suit the current needs of the organism; (vi) coordination patterns adapt to changing internal and external contingencies; (vii) depending on a balance between competitive and cooperative processes, learning may take the form of abrupt transitions from one coordinated pattern to another; and (viii) the system may remain in the current pattern of coordination even when conditions change thus exhibiting memory.

The foregoing list contains some of the core aspects of CD reflecting its inherently nonlinear and emergent character. Such phenomena appear so spontaneously and so consistently as to suggest the existence of an underlying lawfulness or regularity that transcends the multitude of differences between different systems and the settings in which they can be observed [41,45].

Coordination achieves its pinnacle in the vast array of cells and connections called the human brain, and in the collection of human beings we call society [41,45]. How social interactions form and change in complex systems and contexts is of great interest to many disciplines, particularly psychology, biology, physics, economics and the social sciences. The primary focus of the present article is to review recent work investigating the coordination dynamics of individuals interacting with each other in real time. At the core of all personal relationships is how the other becomes intertwined with the self. *Social coordination* is the tendency of two or more individuals to coordinate their ongoing actions with each other based on mutual information exchange. *Social Coordination Dynamics* (abbreviated SCD) is a theoretical-empirical framework that investigates the behavioral and neural dynamics of bond formation between individuals, operationalized in terms of how they spontaneously synchronize their behavioral and neural patterns [82,99].

Synchronization is a form of spontaneous pattern formation that operates according to general principles of self-organization described by nonlinear dynamics [25,26,72]. Following on Huygens's analysis of two clocks synchronizing on a wall, many studies have framed the problem of mutual synchronization in terms of a network of oscillators each of whose individual behavior is altered by nearest neighbor interaction [5,7,30,57,106,107]. Under that framework synchronization has been observed among very different entities in a broad range of physical, biological and social systems. Human brains (and behavior) have proven no exception to these principles [19,27,41,50,51,89]. Experiments have revealed that humans exchange information – whether uni- or multimodal in nature – to spontaneously adopt and switch coordination patterns (e. g. [37,53,58]).

The validity of the measures and constructs from coordination dynamics are worth mentioning because they speak to the appropriateness of a dynamical framework for investigating social situations. Whereas it is easy to justify the physical existence of linkages between components in the coordinated behavior of a single entity, no such linkage typically exists between people. *Social coordination occurs via information exchange*, typically through vision, touch and sound. Emotional interactions may also be involved. A natural measure that describes this informational exchange is the relative phase between coordinating behaviors. The relative phase is an informational variable whose dynamics and has been shown to capture quantitatively coordinated patterns of brain and behavior among different kinds of components, events and processes (see ► Coordination Dynamics [26,40,41]). For coupled rhythmic behaviors, the relative phase dynamics is often adequate not only for uncovering basic mechanisms underlying synergy formation and behavioral change but also the strength and directionality of influences during social interaction ([82,99]; see also [62,87]) for recent reviews.

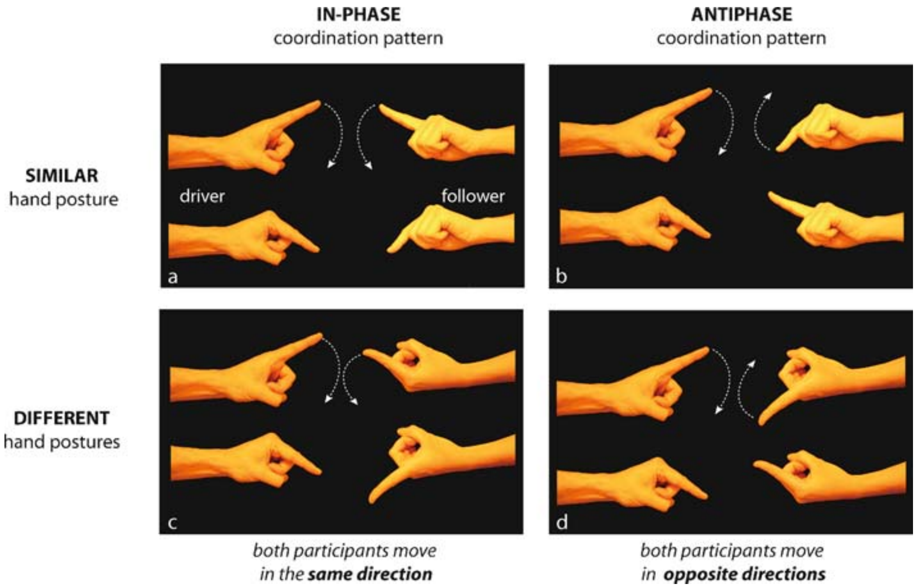
Intentional Interpersonal Coordination

Among the many phenomena of human social coordination, one that most of us have experienced is the synchronized clapping of an audience. Néda and colleagues [70,71] have investigated why applause occurs in unison, with individual "clappers" sometimes acting as a single synchronized ensemble. Although synchronized clapping may vary little from one situation to another, the mechanisms governing the phenomenon are nuanced and context-dependent, even within the same audience. An illustration of this context-dependence comes from the world-famous *New Year's Concert* given every year by the Vienna Philarmonic Orchestra in Austria. Traditionally the concert ends with the Radetzky March by Johann Strauss Jr. This piece of classical music is performed in quite an unusual way. For instance, the conductor leads not only the orchestra but also the audience. Upon a visual cue from the maestro, the audience claps in synchrony with the music. The collective clapping is synchronized both with the music and the visual signals given by the conductor. The reader who is not really into classical music might prefer the example of the song *'We will rock you'* by Queen. Except for a final and unique guitar solo, this song is constituted by a powerful rhythm and a poignant vocal performance by lead singer Freddie Mercury. When this very rhythmic song was performed live, the audience intentionally coordinated its movements with the sound of the drums and the pattern of movements visually provided by the singer. People were therefore intentionally clapping their hands on the first two beats and extending their arms on the third. In the coordination dynamics literature, this is referred to as *intentional sensorimotor coordination* of individuals with external events.

Several studies have employed the sensorimotor coordination tasks to investigate interpersonal coordination dynamics for the case when a person intentionally synchronizes her movements with another by means of visual information exchange see [62] for a review. Following the bimanual and sensorimotor paradigms introduced by Kelso and colleagues [36,37,46,48], Schmidt, Carello and Turvey [88] asked two individuals sitting next to each other to swing their legs in an in-phase or antiphase fashion with respect to the leg movements of the other member of the dyad. As movement frequency was increased (or decreased) by means of an auditory metronome, they found many of the predicted features of nonequilibrium phase transitions [27,47,89]: (i) *differential stability* between the two coordination patterns; (ii) *phase transitions* from the less stable coordination pattern (antiphase) to the more stable one (in-phase); (iii) *critical fluctuations* (i. e. increase in coordination variability) in the vicinity of the transition region and (iv) *hysteresis* (i. e. a sensitivity to the history of the system). All such hallmarks of coordination dynamics (and others such as *critical slowing down* in the vicinity of transitions) have been repeatedly found in a huge number of studies covering various experimental settings. Those settings include, but are not restricted to inter- (e. g. [36,37]) and intra-limb coordination (e. g. [9]) and coordination beween a limb and its uni- (e. g. [3,9,48]) or multi-modal environment [58] to name only a few. The main contribution of Schmidt and colleagues' research was to demonstrate that coordination phenomena found within a person's brain or body, extend to the interactions between people. It is noteworthy that the observed effects extend outside the typical laboratory setting to include coordination phenomena between an individual and an animal as in Lagarde and colleagues' investigation of the coordination dynamics of the horse–rider system [59]. In this unique experiment horses were ridden while walking, trotting and running on a treadmill. The movement dynamics of the horse, the rider and the horse–rider pair were recorded and analyzed revealing that the human–animal dyad exhibits similar coordination dynamics to human interpersonal coordination [59]. In this respect, it cannot be overemphasized that coordination dynamics deals with emergent cooperative effects across very different coordinating elements from neurons to muscles to limbs to people and across the animal–environment divide ([41,42]; for an excellent discussion, see Turvey [100] and commentaries in Vallacher and Nowak [101]). Both the 'intrinsic dynamics' of the individual elements and the nature of the coupling between different elements must be identified for a full account of the phenomena observed.

Several experiments by Schmidt and co-workers, as well as by other groups, have explored the effects on interpersonal coordination of variables such as the manipulation of objects (e. g. hand-held pendulums) or visual surroundings ([74,86]; see [87] for a recent review). Incorporating both aspects de Rugy and colleagues developed a neuro-mechanical model of visually mediated intentional interpersonal coordination [16]. Their model consists of two cross-coupled neuro-mechanical units, each composed of a neural oscillator driving a wrist-pendulum system moved by a different person. Taken individually, each unit reproduces the natural tendency of the participants to freely oscillate close to resonance frequency. When cross-coupled through the vision of movements of the other individual, each person entrains the other as they adopt a common frequency influenced by their own mechanical properties. Although important, neuromechanical properties are not the only factors that determine the stability of coordination patterns between individuals: attentional load and egocentric constraints also influence interpersonal coordination dynamics [96,97].

A series of experiments has investigated whether the motoric and perceptual constraints that shape the dynamics of inter- and intra-limb coordination play a similar role in the coordination between people (e. g. [12,13,53,58,66,78]). In intrapersonal bimanual coordination the preference for co-activation of homologous muscles appears to be mediated by general principles of symmetry in neural organization such as reciprocal connectivity between homologous brain areas. In a study by Oullier and colleagues [76] investigating the relative role of visual/directional and motor (a) symmetries in interpersonal coordination, two participants made index finger flexions while seated facing each other. One acted as a driver (D) by synchronizing to a metronome that systematically increased in rate. The second participant, or follower (F), was required to coordinate finger movements with D via visual coupling only. F participated in four conditions (Fig. 1) determined by a combination of coordination pattern (in-phase or antiphase) and hand posture (supination or pronation). The relative phase requirement was defined by the spatial configuration (i. e. the position of the endpoint of the finger). In this way, co-activation of homologous muscles (finger flexion by F and D) produced both an in-phase and antiphase relationship between the effector endpoints depending on the experimental condition. If purely directional constraints [92] determine the stability of interpersonal coordination, and D functions only as a generic rhythmic stimulus, perceptual antiphase coordination should display decreased stability regardless of the relative hand position of the participants. Contrary to this hypothesis, a strong role was found for interpersonal homologous muscle co-activation. Coordination between individuals was most stable when they were activating similar muscle groups such that co-flexion was always more stable regardless of the resulting spatial pattern. Directional constraints played only a modulatory role. These initial results are at odds with the concept of social coordination as a form of simple perceptual-motor coupling. Rather, it appears that perception of homologous muscular activation acts as a constraint on coordinative stability, creating a "functional homology" to bimanual coordination. Thus, social coordination may be differentiated from



Social Coordination, from the Perspective of Coordination Dynamics, Figure 1
Participants show a preference for homologous muscular activation, irrespective of the visuospatial congruency of their movement. Each column describes a possible configuration of interpersonal coordination. The left participant is paced with a metronome whose frequency increases (driver). Oullier and colleagues [76] studied the frequency at which the right subject (follower) loses stability in each condition. In columns a and b, both subjects are in the same hand position. The pattern a (both flex then both extend: in-phase coordination) is more stable than the pattern b (when one extends the other flexes: anti-phase coordination). In columns c and d, the participants adopt a different hand posture. The pattern d (both flex then both extend: anti-phase coordination) is more stable than the pattern c (where both subjects move in the same direction). These results suggest that coordinative stability is not purely governed by visuospatial congruency (cf. [67]). Rather, the embodiment of the other's movement leads the follower to adopt an anatomically homologous movement. This pattern of behavior is unique to the fact that the follower and the entity with which he/she coordinates are both humans [76]

simple perceptual-motor coupling by virtue of the biological and functional relevance provided when viewing another person.

Although all these studies have employed comparable experimental settings and the common theoretical framework of coordination dynamics with the aim of better understanding intentional interpersonal coordination, it is not yet clear whether spontaneous mutual entrainment actually occurs in a true two-way interaction, or whether one individual simply acts as a pacing stimulus or 'driver' for the other (e. g. [48]). A similar concern can be raised regarding the behavior of the audience during the Radetzky March at the New Year's Concert in Vienna. It seems unlikely that audience members spontaneously synchronized with each other while music was played, since their primary intent was to respond by clapping in rhythm with the music and with the visual cues coming from the stage. This process has been well described in human movement (neuro)science and coordination dynamics and occurs when an individual intentionally coordinates his movements with external physical stimuli [3,41,48]. A sensorimotor interpretation of audience participation is strengthened by results of an experiment in which the auditory metronome used to pace the interpersonal coordination was silenced at times [76]. The study revealed that the presence of an external pacing stimulus (an auditory metronome in that case) actually reduced interpersonal coordinative stability regardless of the adopted directional or muscular pattern adopted. Oullier and colleagues [76] provided evidence for stronger mutual entrainment when no external information could perturb the dyadic interactions, analogous to what the music and the conductor would do during a concert. Hence, from an experimental perspective, this phenomenon is not social interaction per se but rather sensorimotor coordination to an external event. In the case of the "observed" audience clapping in unison *during* the performance, one could argue that the audience is constituted by a collection of individuals coordinating mainly with the music and the conductor with little contribution from neighbor-to-neighbor interactions, (A similar concern can be raised in the study by Schmidt and co-workers [88] as participants were instructed to intentionally coordinate with each other *and* with an auditory metronome. Hence, one participant could serve as a visual metronome to the other (and reciprocally) and/or the phase transitions observed could either be interpersonal in nature (from interpersonal antiphase to in-phase) or from syncopation to synchronization as in a single individual coordinating with an auditory metronome (cf. [48]).

A different scenario, however, is characteristic of the end of the performance, when the audience expresses its approval of the orchestra and conductor through applause. At this moment each person applauds according to her preferred/intrinsic pace with no driving stimuli – whether visual or auditory – coming from the stage. In spite of the absence of pacing information, the audience quickly and spontaneously entrains to a common rhythm such that everyone is clapping in unison. Note that at this moment, the only information that can alter an individual's behavior is the sound (and possibly the vision) of the movements made by their neighbors [70,71]. Thus, we have units involved in individual rhythmic behaviors communicating via, at least, one means of information exchange. According to Winfree [108], this is a minimum requirement for self-organized spontaneous synchronization to emerge (see also [41]). In that case, any collective pattern that emerges is more likely to be unintentional compared to situations where the audience follows the conductor and the music.

Issues in Quantifying Spontaneous Interpersonal Coordination

An abundant literature exists addressing unintentional interpersonal coordination in experimental paradigms ranging from people swinging pendulums [86], dancing [33], walking [102] or rocking chairs [84] to performing joint Fitts' tasks [69], talking to each other [83,91] or even boxing [60]. However, many questions remain regarding the nature of the behavioral and neural processes mediating the formation and dissolution of unintended synchronous behavior between individuals and how such processes may be quantified [2,55].

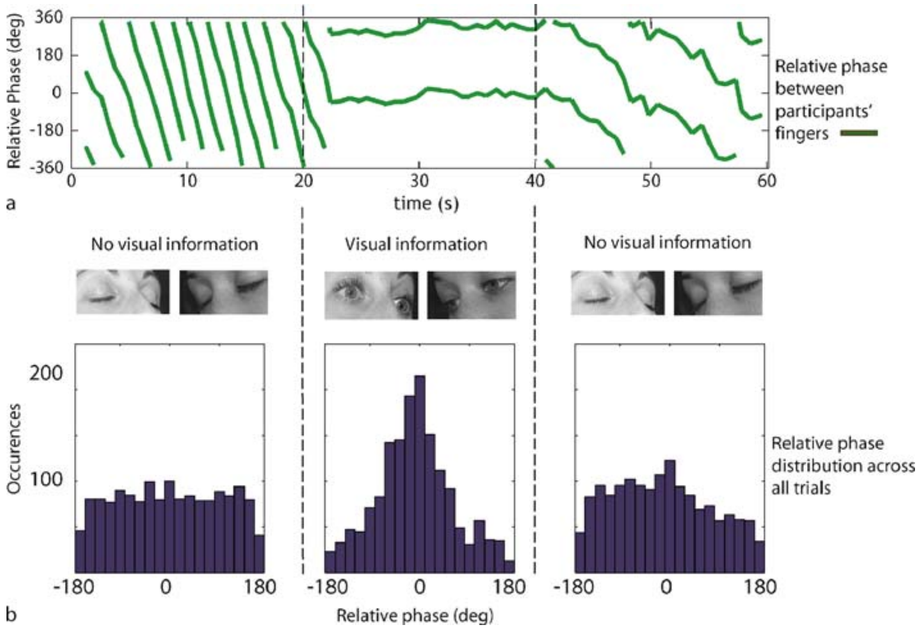
Oullier and colleagues [82] have identified three major problems in investigating spontaneous synchronization in social settings. First, even when the source and nature of the coupling has been identified, it is difficult to manipulate experimentally relevant variables such as the coupling strength (e. g. [71]). Almost by definition, spontaneous behavior is not externally goal directed or explicitly controlled. Most of the results reporting unintentional synchronization in humans are based on observation and categorization methods that rely primarily on the experimenter's appreciation of a given exemplar behavior rather than a quantitative measure of coupling and individual behavior (e. g. [4,14]).

A second problem is the challenge of complexity, both in terms of the large number of units to analyze (e. g. thousands of pairs of clapping hands [71]) and the complexity of the behavior itself (e. g. mother-infant synchronization [14]). Such compositional and behavioral complexity has hindered experimental attempts to record and quantify both the individual and social dynamics. Even the reduction in dimensional complexity afforded in coordinated behavior can only go so far in elucidating the relationship between group behavior and the individual units of which it is composed.

A third problem comes from the possibility that any change in a person's behavior induced by interacting with another may persist even after the encounter is over. So far, there has been very little precise quantification of the mutual influence people have on each other's behavior a posteriori, i. e. how individual behavior is affected after the social encounter when people no longer exchange information (but see Sect. "Social Memory and the Dependence on Initial Conditions").

Human Spontaneous Synchronization

In behavioral experiments that revealed spontaneous interpersonal synchronization, Oullier and colleagues [77, 80,82] explored coordinative patterns that emerge only as a function of visual information exchange. The main hypothesis was that even without instructions to do so, spontaneous synchronization between partners would occur as soon as they coupled visually while moving in front of each other. On the other hand, spontaneous interpersonal coordination should disappear whenever exchange of information is no longer possible. In Oullier et al.'s behavioral experiments, pairs of participants executed movements while in full view (or not) of each other's ongoing actions as well as their own [77,80,82]. Each member of the dyad executed movements at their own preferred frequency and amplitude without any external pacing from a metronome or any other sort. Movements were required to be as smooth and continuous as possible throughout an experimental trial. What is important here is that participants were not given any instructions regarding the way to move with respect to each other. The experimental protocol consisted of participants moving with no vision of the other's movements before being allowed to see their own actions at the same time as they saw the other person's. Finally, visual information was removed again. Experi-



Social Coordination, from the Perspective of Coordination Dynamics, Figure 2

Relative phase between the movements of two individuals. a This panel illustrates the evolution of the relative phase as a function of time in a representative trial of the SCD paradigm. *Left column* No visual information exchange, each individual moves independently, movements are uncoordinated. *Middle column* as soon as people exchange visual information, they spontaneously couple. Their relative phase is therefore close to 0° . *Right column* When visual information is removed, they are no longer synchronized. b This panel represents distributions of relative phase for all the subjects and all the trials (adapted from [82])

mental trials were therefore equally partitioned into three contiguous segments each of equal duration within which both subjects either were allowed to exchange information with each other or not. When visual information was available, participants looked at each other's finger motion and were also able to see their own finger [77,80,82].

In SCD, following theories of cooperative phenomena in open systems [25,26] a central idea is that the behavior of a complex dyadic system may be captured by the value of a low-dimensional collective variable known as the *order parameter*. In the vicinity of critical points, emergent behavior is governed by the dynamics of this collective variable e. g. [25,41]. In experimental cases the order parameters are not known in advance but have to be discovered. For the situation of social coordination as in many other cases treated by CD, an appropriate order parameter describing the system dynamics is the relative phase ϕ between the movements of each member of the pair [80,82]. The relative phase measure allows for a reduction of a potentially very high dimensional system (e. g. where one has to consider, among other components, the neurons, joints and muscles of both individuals) as it captures the macroscopic spatio-temporal behavioral pattern (see Fig. 2). Even at an overt behavioral level, four degrees of freedom (position and velocity of each component) may be compressed onto a single relative phase value that summarizes the organization of the dyadic system. Quantitative evaluation of spontaneous synchrony is also provided by the FFT power spectrum overlap between the movements of each person. The spectrum overlap measures the percentage of movement frequencies common to both partners in a pair [82]. Defined as the area of intersection between each participant's normalized spectral plots, it serves an indicator of the strength of the frequency entrainment between the two participants (see Fig. 4).

When no visual exchange was allowed, each subject produced movements independently at their own frequency. As a result, the relative phase ϕ between the subjects' finger motions exhibited phase wrapping (Fig. 2, left column). However, following a simple auditory cue to open their eyes, subjects spontaneously adopted in-phase motion, ϕ stabilizing around 0° (Fig. 2, middle column). On a signal to close the eyes again, the individual movement frequencies diverged and ϕ fell back into phase wrapping (Fig. 2, right column). These initial results were corroborated by a subsequent more extensive study, in which the order of the vision and no-vision segments was changed. Once again, spontaneous synchronization emerged as soon as vision of the other's movements was allowed [82]. Overall, results reveal that with visual information exchange, participants tend to mutually couple at a common phase and frequency, whereas in the absence of vision, participants' movement trajectories diverge and behave independently. Such emergent mutual coupling is truly a result of spontaneous social interaction and may be distinguished from previous dyadic studies in which one person may simply be intentionally tracking (or driving) the other [16,77,88,97] or maintaining their own rhythm [86].

Why does spontaneous interpersonal coordination occur at all? Compelling examples stretching from human evolution through religious ritual and sports to political, war and economical strategy suggest that *keeping together in time* is one of the most powerful ways to create and sustain communities and communication [65]. Moreover, *not* moving in synchrony may be too costly for the dyad see, (e. g., [56]).

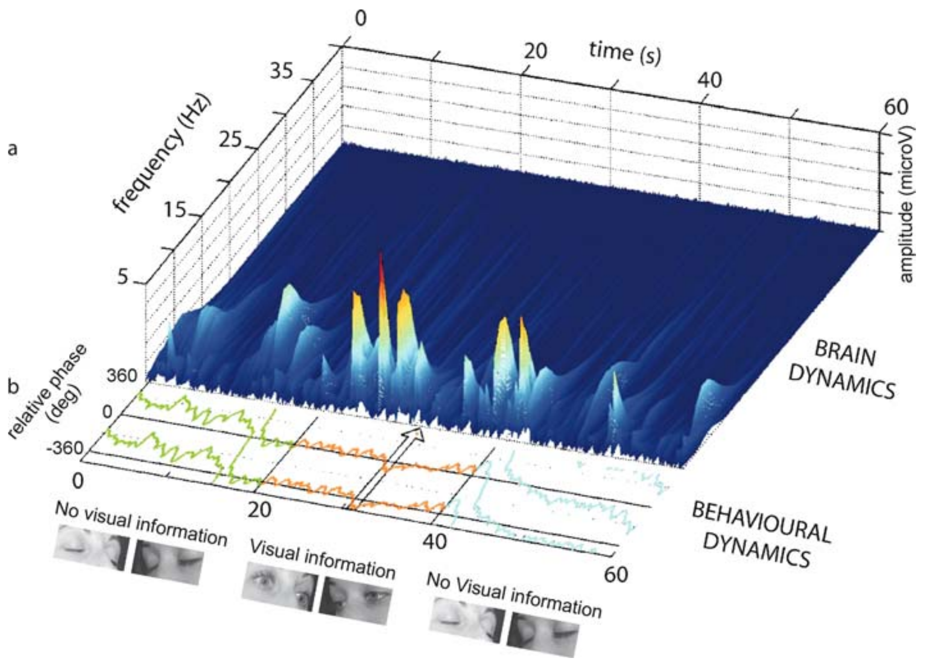
In order to better understand which features of visual information exchange may facilitate spontaneous social coordination one has to bear in mind that human movements can be unintentionally affected by the vision of an object oscillating in the environment. This is illustrated by experiments using the moving-room paradigm in which the walls of the room move but not the floor (e. g. [61,75,78]). Body sway of the observer's couples in time spontaneously with small oscillatory motions of the room. In addition, experimental data show that the mere observation of the movements of another person interferes with one's execution of a similar action [54]. Interestingly, such interference is less noticeable when the movements observed are not generated by humans [15]. In the latter work, one of the members of the dyad was replaced by a computer-generated moving hand, the trajectory of which was driven either by a sinusoidal function or a pre-recorded real finger trajectory. The stimulus movement frequency in the study by de Guzman and coworkers [15] was fixed at either 10% below or 10% above the subject's self-paced rate as determined at the start of the experiment. Results revealed that the human–avatar coordination was strongest when the latter was an image of a hand driven by real movement data. The weakest coupling occurred when the visual stimulus followed a sinusoidal trajectory. Unlike the interpersonal situation [82], spontaneous synchronization was not found for all trials and, when it happened, was supported by a significantly lower frequency overlap [15]. One may invoke a one-way coupling to explain these findings, since the motion of the computer generated hand could not be influenced by the movement of the participant. Taken together, the foregoing results support the hypothesis that biological relevance in general, and biological motion in particular – including its natural variation– play a key role in social coordination.

Shared Behavioral and Neural Social Coordination Dynamics

One explanation for the emergence of spontaneous social coordination may be found at the neurophysiological level. For instance, some areas of the brain are known to be associated with the perception (but not the execution) of biological motion including the posterior superior temporal sulcus (abbreviated STS) [1,23,24,31]. STS is known to be a major source of visual information for the so-called *human mirror system* (abbreviated HMS) [85]. Originally identified in monkeys, *mirror neurons* are (sensori)motor neurons discharging both when one performs a given action and sees the same action performed by someone else. They have been identified primarily in the ventral premotor cortex and the rostral region of the inferior parietal lobule [20]. The HMS constitutes a neural mechanism that is automatically activated by the sight of somebody else's actions, even when the observer does not make overt movements. The main idea is that during observation the HMS provides a simulation of the actions of other people potentially providing a basis for understanding the intentions of others [31].

Since the foregoing behavioral experiments allow participants to both produce and observe movement at the same time it seems possible that the HMS is at least partially involved in the spontaneous coordination observed. In order to investigate this question, Tognoli and colleagues [99] recorded brain activity of each member of the dyad using a specially designed dual-electroencephalography (EEG) system. Each participant wore a 60-electrode EEG-cap that enabled simultaneous recording of their brains to accompany kinematic measurements of their behavior.

To grasp the significance of the work by Tognoli and colleagues, we need to revert to earlier studies conducted within the framework of Coordination Dynamics have employed instabilities in coordination as a means to uncover the link between the dynamics of behavior and the dynamics of the brain [39,42], with the goal of relating levels by virtue of their *shared dynamical properties* (e. g. [19,39,49,50,52]). In such research, the high temporal resolution of electroencephalography (EEG) and magnetoencephalography (MEG) was exploited to quantify the relationship between the large scale neural dynamics emerging from billions of interconnected neurons and the behavioral dynamics revealed in experiments on coordi-



Social Coordination, from the Perspective of Coordination Dynamics, Figure 3

Relation between Phi_2 and social coordination. a Time-frequency spectrum from electrode CP4 (located over parietal brain regions) from a single trial. Phi_2 is low before and after vision but increases during vision. b Corresponding relative phase between finger movements. Synchronized in-phase behavior is observed during visual contact. Notice the temporary disengagement of the rhythm when coordination is lost briefly (adapted from [99])

nation [41]. Observed features of the dynamics expressed at both levels of description such as multistability and phase transitions (i. e. the spontaneous switch from one pattern to another due to loss of stability), were taken as evidence that principles of self-organization govern pattern formation in both brain and behavior [26,41]. Of particular initial interest was the identification of qualitative changes in the pattern of neural activity that occurred simultaneously with transitions between behavioral patterns [19,49,50,63,104]. On the basis of this work, an exciting hypothesis is that the transitions from uncoordinated to spontaneous coordination observed in the SCD paradigm may be accompanied by similar events at the brain level.

In an effort to shed new light on how social processes are integrated in the brain, Tognoli and colleagues [99] identified several neural mechanisms or *neuromarkers* that appear and disappear with the emergence and dissolution of coordinated behavior between two people. Interestingly, these social neuromarkers consist of brain rhythms in the 10 Hz frequency range located over right centro-parietal areas of the cerebral cortex. In particular, a social brain rhythm termed the *Phi Complex* consists of two components: the first, Phi_1, increases during independent behavior i. e. before information exchange between members of the dyad. When subjects saw each other's finger movements and coordinated together, Phi_1 disappeared and Phi_2, a different rhythm within the same frequency band appeared (Fig. 3) [99].

In a subsequent study, Tognoli and colleagues [98] explored the dynamics of the *Phi Complex* by instructing participants to *intentionally* synchronize when visual information exchange was allowed. In this case, participants interact to accomplish a shared goal. Again, Phi_1 appeared during uncoordinated behavior and Phi_2 when social coordination occurred. Analysis of dyads who participated in both experiments [98,99] revealed that the amplitude of Phi_2 was higher during intentional than spontaneous coordination. Thus, Phi_2 appears to be a neural signature of social coordination whether it emerges spontaneously or not.

The cortical location of the *Phi Complex* appears to be consistent with neuro-anatomical sources within the human mirror system. One of the conclusions drawn by Tognoli and colleagues is that Phi_1 might have an inhibiting role on the mirror system. Previous claims by Brass and Heyes [8] have argued that the mirror system is always active by default and thus must be inhibited in non-social contexts. Hence Phi_1 could inhibit Phi_2, the latter being seen as a facilitator of social coordination that participates in information exchange between the motor cortex and the mirror system [98,99].

In summary, experiments using dual high density electrode arrays to record and measure brain activity from two persons in conjunction with motion capture technology, have allowed an exploration of shared behavioral and neural social coordination dynamics [98,99]. Transitions from uncoordinated dyadic behavior to interpersonal synchronization have been demonstrated to accompany the emergence of a new brain rhythm – the Phi complex – located in the human mirror system. Such work suggests that SCD may serve as a novel framework for identifying behavioral and neural signatures in reciprocal interactions and allows for a more dynamical approach to the study of the mirror neuron system.

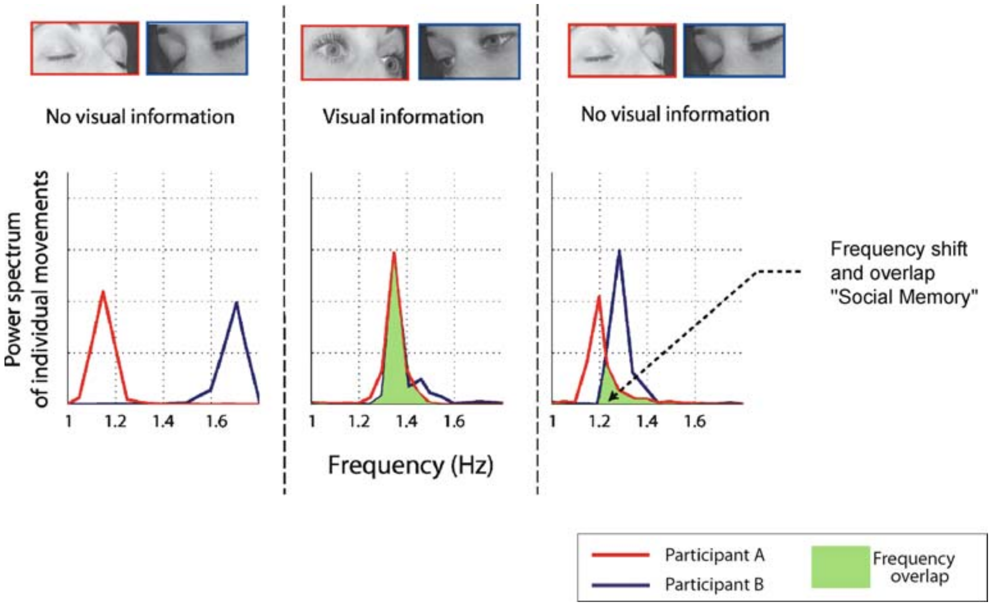
Social Memory and the Dependence on Initial Conditions

At first blush, the emergence of spontaneous coordination between individuals [77,80,82,99] might be seen as an instantiation of mutual entrainment that entails nothing more than a couple of oscillators and a medium of information exchange [41,108]. In such generic cases, once the coupling is removed, each oscillator should return to its own intrinsic frequency, that is, any influence of the interaction should disappear. However, the situation between two people is different (see Fig. 4). Theoretically, in a typical coupled clocks scenario, there should be no difference between the movement periods of the 'clocks' before or after coupling-induced synchronization. However, a serendipitous experimental finding [82] was the consistent and persistent influence of the social interaction on subsequent rhythmic behavior despite the absence of information exchange between the pair (Fig. 4). This remnant of a prior social interaction may qualify as a kind of *social memory* [82].

Social memory is thought to play an important role in human actions, and, to a larger extent, on the way we live [32]. In the context of SCD, social memory implies that the intrinsic parameters of the individual components have been altered by virtue of the social interaction. Mathematically [42], one may represent the situation before the interaction as follows:

$$\begin{aligned} \ddot{x}_1 + i_1(x_1, \dot{x}_1, a_1) + \omega_1^2 x_1 &= 0 \\ \ddot{x}_2 + i_2(x_2, \dot{x}_2, a_2) + \omega_2^2 x_2 &= 0\,. \end{aligned} \qquad (1)$$

Where x_1 and x_2 represent the coordinating components,



Social Coordination, from the Perspective of Coordination Dynamics, Figure 4

Evidence for social memory. Illustrated is an example of frequency overlap between the movements of both subjects in a representative trial of the SCD paradigm. *Left column* no vision of each other's movements; each individual moves at their own intrinsic frequency so there is no frequency overlap. *Middle column* visual information is exchanged between participants, causing spontaneous synchronization to occur at a common frequency (and phase, see Fig. 2). *Right column* visual information is no longer exchanged but individuals do not revert back to their initial intrinsic frequency. This remnant of frequency overlap as a result of prior social interaction suggests a kind of 'social memory' (adapted from [82])

$i_{1,2}$, $a_{1,2}$ and $\omega_{1,2}$ refer to individually chosen intrinsic parameters such as the chosen frequency and amplitude.

During the interaction, the system is visually coupled, F_1 and F_2 representing a coupling function such as the well-known HKB-coupling [27,44]:

$$\begin{aligned} \ddot{x}_1 + I_1(x_1, \dot{x}_1, A_1) + \omega_1^2 x_1 &= \mathrm{F}_1(x_1, \dot{x}_1, x_2, \dot{x}_2) \\ \ddot{x}_2 + I_2(x_2, \dot{x}_2, A_2) + \omega_2^2 x_2 &= \mathrm{F}_2(x_2, \dot{x}_2, x_1, \dot{x}_1)\,. \end{aligned} \tag{2}$$

Now notice that the interactive context has formed a coupling (the right hand side of Eq. (2)) but also led of a modification of the individual component parameters, I and A (on the left hand side of Eq. (2)). One may say that the boundary conditions of Eq. (1) have been altered by the social interaction.

After the interaction, the coupling function disappears (F_1 and F_2 terms on the right hand side are zero) and the system is "uncoupled" (cf. Fig. 1):

$$\begin{aligned} \ddot{x}_1 + I_1(x_1, \dot{x}_1, A_1) + \omega_1^2 x_1 &= 0 \\ \ddot{x}_2 + I_2(x_2, \dot{x}_2, A_2) + \omega_2^2 x_2 &= 0\,. \end{aligned} \tag{3}$$

However, notice in Eq. (3) the individual intrinsic parameters of the system which were modified by the interaction are still in place. Though uncoupled, the individual components are still affected by the interaction. How this internalization process occurs remains open to empirical investigation.

A benefit of the SCD paradigm is that one is able to quantify the strength and persistence of prior social influences on an individual's behavior. The finding that the modification of the neural network depends on which modality is engaged during the mutual encounter suggests that additional cortical areas may have been recruited and included into the initial global neural assembly due to social context [32]. However, beyond the Phi complex, and perhaps due to inherent limitations in spatial resolution, examination of the dual EEG data showed no evidence of further cortical engagement. Another possibility (in line with the foregoing mathematical analysis) is that the connectivity and dynamics of the initial network is modified by social interaction, and the new organization retained after the interaction is over. Recent evidence in support of this hypothesis suggests that two people engaging in a common task share a representation of each other's movement dynamics, including trajectory amplitude and frequency [6,17]. Such a (shared) representation may persist when vision is removed, i. e. when information exchange is no longer possible [21]. Moreover, representations at the neural level have been shown to be highly flexible and context-dependent [34,35], influenced both by environmental [105] and task demands [79].

The extent and duration of the carryover or remnant effects observed in behavioral experiments may reflect many factors, including the strength of the bond that is formed between people, place in the social hierarchy, the willingness of each participant to cooperate, gender differences, personality characteristics and the significance each participant attaches to the social encounter [32]. An additional finding from our work favors a motor contribution to social memory as well: the persistence effect was found to be independent of the duration of movement that followed the social encounter [82]. This hypothesis is strengthened by results showing that observation of another person performing movements generates a kinematically specific memory of the observed motions in primary motor cortex [95].

The systematic directionality effect observed in the SCD paradigm is revealing also [82]: the extent to which one member of the dyad is influenced by the other was shown to depend on initial conditions. Obviously for synchronization to occur, the person moving with the lowest/highest intrinsic movement frequency must speed up/slow down during information exchange. A surprising result is that the difference between the initial and the final intrinsic movement frequencies (vision absent) was always greater for the person starting with the higher compared to the lower movement frequency [82]. The extent to which initial, so-called 'intrinsic dynamics' determine behavior after the social encounter is over may be of great interest to understanding social interactions in more complex settings where hierarchical relations are involved.

An important problem in human social behavior concerns understanding the degree to which an individual influences the actions of a group (e. g. peer group, family, class) he/she is in. Due to several factors (personality, situational), a person (the leader) may affect the behavior of others more than the others affect her or him. The concept of leadership is commonly associated with interactions taking place in hierarchical settings such as typical organizations, but is actually broader than that. Strength of behavioral influence is overlooked because behavioral interactions have not been systematically studied. Contemporary complex systems approaches (e. g. [28]) view the formation of leader-follower roles as interactive and emergent but in so doing may have undermined the significance of individual dimensions. The approach of SCD is rather to ask: *What makes two people behave independently and what makes them behave as a unit?* The paradigm of social coordination dynamics exploits inherent asymmetry between two people during behavioral interactions and gauges, e. g. using directional coupling measures, which of the two has a stronger influence.

Future Directions

Human beings are social by nature, and interactions with others represent a substantial portion of their many daily activities. A common and well described consequence of interpersonal activity is that an individual's behavior, whether intentional or not, is modified by interactions with others [32]. Alterations of individual and collective behaviors range from imitation and mimicry to spontaneous synchronization, and have been observed in groups varying in size from dyads to thousands of individuals e. g. [4,68].

Social Coordination Dynamics investigates how the natural (uninstructed) social influence of one person on another evolves in real time and has led to a number of new findings. The first is that humans immediately and spontaneously coordinate their actions with each other when provided vision of the movements of the other together with their own. The second is that a specific brain rhythm underlies social coordination. Transitions from individual to coordinated social behavior are observed at both behavioral and brain levels.

The third finding is that an individual's intrinsic behavior is altered by social interaction: the effect of the previous social encounter persists when vision of the other's movements is no longer available. A fourth and final finding is that social coordination is affected by initial conditions, enabling one to predict which individual is most affected by the social encounter.

Insights into elementary forms of social interaction have been obtained by applying the concepts, methods and tools of coordination dynamics. A notable feature of coordination dynamics is its ability to uncover mechanisms and principles common to different kinds of complex systems at different levels of observation and to relate them by virtue of shared behavioral and neural dynamics [41]. SCD and its dynamical measures have proven to provide adequate quantification of the spontaneous coupling between individuals, the transition to loss of entrainment and the effect of the social encounter at both behavioral and brain levels. The same basic patterns of coordinated behavior and pattern dynamics (multistability, critical fluctuations accompanied by a temporary loss of stability, phase transitions, hysteresis and critical slowing down) have been observed within an individual, between an individual and the environment, and between individuals. In this respect SCD complements recent developments in social cognitive neuroscience, behavioral economics, game theory, socio-economics and neuroeconomics (e. g. [10,11,18,73,81,94,103]).

The field of *social neuroeconomics* serves to illustrate the benefits of considering SCD in contexts other than interpersonal sensorimotor interaction. Social neuroeconomics investigates the neural correlates of economical decision making [18]. One particular feature of this nascent field is that decision making processes are always studied in a body- and movement-independent fashion. Why is that? After all, from the very first months of life, individuals live vicariously through one another adopting, if only temporarily, a similar posture or tempo during interactions with a peer, or yawning [4,64,90]. As Henry Greely [22] recently reminded the readership of Science Magazine *"Human society is the society of human brains. Of course those brains are encased in, affected by, and dependent on the rest of the body, but our most important interactions are with other people's brains, as manifested through their bodies."* Although this statement sounds like common sense, thus far the coordination dynamics between bodies has remained unexplored in the field of social neuroeconomics. Yet how many times have we experienced the feeling that trusting someone will be difficult even before talking to them? Whether it was the way she moved or some other factor, body-related cues play a key role in modulating economic decisions (e. g. [93]). A scientific approach to "body language" might aim to understand how perceived actions of others affect the cognitive and emotional processes involved in economical decision-making. For instance, a finding such as the Phi Complex – especially the modulation of Phi$_2$ when individuals intentionally coordinate [98,99] – could turn to be crucial to better competition–cooperation mechanisms underlying decision in economic contexts such as public coordination games [29]. In sum, as a conceptual framework that encompasses the dynamics of both neural and behavioral levels, SCD promises to bridge the gaps between levels of analysis [41,81] and clear a path for new multi-level, interdisciplinary investigations of social interactions. Like synchronization itself, the function of SCD is to facilitate communication across heretofore unrelated fields.

Acknowledgments

The authors wish to thank their collaborators at the *Human Brain and Behavior Laboratory* (Center for Complex Systems and Brain Sciences) who co-authored the articles on social coordination dynamics reported in this article: Drs. Gonzalo C. de Guzman (*Florida Atlantic University*), Kelly J. Jantzen (*Washington Western University*), Julien Lagarde (*Université de Montpellier I*), Cyrille Magne (*Middle Tenesse State University*) and Emmanuelle Tognoli (*Florida Atlantic University*). Amélie L. and Audrey O. should also be thanked for letting the authors use their beautiful hands and eyes in all figures. We wish to thank the following institutions for supporting our research and the writing of this article. The *US National Institutes of Health* (NIMH Innovation Grant MH 42900 and MH080838), the Pierre de Fermat Chair of the Région Midi-Pyrénées (France) to J.A.S.K, and the *Centre National de la Recherche Scientifique* (Programme CNRS Neuroinformatique to O.O.). Correspondence should be addressed to: Olivier Oullier (olivier@oullier.fr) or J. A. Scott Kelso (kelso@ccs.fau.edu). For additional information please visit the following websites: www.oullier.fr and www.ccs.fau.edu/section_links/HBBLv2/index.html

Bibliography

Primary Literature

1. Allison T, Puce A, McCarthy G (2000) Social perception from visual cues: Role of the STS region. Trends Cogn Sci 4:267–278
2. Balaban E (2004) Neurobiology - Why voles stick together. Nature 429:711–712
3. Bardy BG, Oullier O, Bootsma RJ, Stoffregen TA (2002) Dynamics of human postural transitions. J Exp Psychol Hum Percept Perform 28:499–514
4. Barsalou LW, Niedenthal PM, Barbey AK, Ruppert JA (2003) Social embodiment. Psychol Learn Motiv 43:43–92
5. Bennett M, Schatz MF, Rockwood H, Wiesenfeld K (2002) Huygens's clocks. Proc Roy Soc Lond A Math Physic Engineer Sci 458:563–579
6. Bosbach S, Cole J, Prinz W, Knoblich G (2005) Inferring another's expectation from action: The role of peripheral sensation. Nat Neurosci 8:1295–1297
7. Bottani S (1996) Synchronization of integrate and fire oscillators with global coupling. Phys Rev e 54:2334–2350
8. Brass M, Heyes C (2005) Imitation: is cognitive neuroscience solving the correspondence problem? Trends Cogn Sci 9:489–495
9. Buchanan JJ, Kelso JA (1993) Posturally induced transitions in rhythmic multijoint limb movements. Exp Brain Res 94:131–142
10. Camerer CF (2003) Behavioral game theory: Experiments in strategic interaction. Princeton University Press, Princeton
11. Camerer CF, Loewenstein G, Prelec D (2005) Neuroeconomics: How neuroscience can inform economics. J Econ Lit XLIII:9–64
12. Carson RG (2004) Governing coordination. Why do muscles matter? In: Jirsa VK, Kelso JAS (eds) Coordination dynamics: Issues and trends. Springer, Berlin, pp 141–154
13. Carson RG, Kelso JAS (2004) Governing coordination: behavioural principles and neural correlates. Exp Brain Res 154:267–274
14. Condon WS, Sander LW (1974) Neonate movement is synchronized with adult speech - Interactional participation and language acquisition. Science 183:99–101
15. de Guzman GC, Tognoli E, Lagarde J, Jantzen KJ, Kelso JAS (2005) Effects of biological relevance of the stimulus in mediating spontaneous visual social coordination. Society Neurosci Program 867(21)
16. de Rugy A, Salesse R, Oullier O, Temprado JJ (2006) A neuromechanical model for interpersonal coordination. Biol Cybern 94:427–443
17. Decety J, Sommerville JA (2003) Shared representations between self and other: A social cognitive neuroscience view. Trends Cogn Sci 7:527–533
18. Fehr E, Camerer CF (2007) Social neuroeconomics: the neural circuitry of social preferences. Trends Cogn Sci 11:419–427
19. Fuchs A, Kelso JAS, Haken H (1992) Phase transitions in the human brain: Spatial mode dynamics. Int J Bifurc Chaos 2:917–939
20. Gallese V, Fadiga L, Fogassi L, Rizzolatti G (1996) Action recognition in the premotor cortex. Brain 119:593–609
21. Goldman MS, Levine JH, Major G, Tank DW, Seung HS (2003) Robust persistent neural activity in a model integrator with multiple hysteretic dendrites per neuron. Cereb Cortex 13:1185–1195
22. Greely H (2007) On neuroethics. Science 318:533
23. Grèzes J, Fonlupt P, Bertenthal B, Delon-Martin C, Segebarth C, Decety J (2001) Does perception of biological motion rely on specific brain regions? Neuroimage 13:775–785
24. Grèzes J, Armony JL, Rowe J, Passingham RE (2003) Activations related to mirror and canonical neurons in the human brain: An fMRI study. Neuroimage 18:928–937
25. Haken H (1983) Synergetics: An introduction. Springer, Berlin
26. Haken H (1996) Principles of brain functioning: A synergetic approach to brain activity, behavior and cognition. Springer, Berlin
27. Haken H, Kelso JAS, Bunz H (1985) A theoretical-model of phase-transitions in human hand movements. Biol Cybern 51:347–356
28. Hazy JK, Goldstein JA, Lichtenstein BB (2007) Complex systems leadership theory: New perspectives from complexity science on social and organizational effectiveness. ISCE Publishing Company, Mansfield
29. Hichri W, Kirman AP (2007) The emergence of coordination in public good games. Europ J Phys B 55:149–159
30. Hugenii C (1673) Horologium oscillatorium. Apud F. Muguet, Paris
31. Iacoboni M, Molnar-Szakacs I, Gallese V, Buccino G, Mazziotta JC, Rizzolatti G (2005) Grasping the intentions of others with one's own mirror neuron system. PLoS Biol 3:529–535
32. Insel TR, Fernald RD (2004) How the brain processes social information: Searching for the social brain. Annu Rev Neurosci 27:697–722
33. Issartel J, Marin L, Cadopi M (2007) Unintended interpersonal co-ordination: can we march to the beat of our own drum? Neurosci Lett 411:174–179
34. Jantzen KJ, Steinberg FL, Kelso JAS (2004) Brain networks underlying human timing behavior are influenced by prior context. Proc Nat Acad Sci USA 101:6815–6820
35. Jantzen KJ, Steinberg FL, Kelso JAS (2005) Functional MRI reveals the existence of modality and coordination-dependent timing networks. Neuroimage 25:1031–1042
36. Kelso JAS (1981) Contrasting perspectives on order and regulation in movement. In: Long J, Baddeley A (eds) Attention and performance IX. Erlbaum, Hillsdale, pp 437–457
37. Kelso JAS (1984) Phase-transitions and critical behavior in human bimanual coordination. Am J Physiol 246:1000–1004

38. Kelso JAS (1991) Behavioral and neural pattern generation: The concept of NBDS. In: Koepchen HP, Huopaniemi T (eds) Cardiorespiratory and motor coordination. Springer, Munich, pp 224–238
39. Kelso JAS (1992) Coordination dynamics of human brain and behavior. Springer Proc Phys 69:223–234
40. Kelso JAS (1994) The informational character of self-organized coordination dynamics. Hum Mov Sci 13:393–413
41. Kelso JAS (1995) Dynamic Patterns: The self-organization of brain and behavior. MIT Press, Cambridge
42. Kelso JAS (2000) Principles of dynamic pattern formation and change for a science of human behavior. In: Bergman R, Cairns RB, Nilsson LG, Nystedt L (eds) Developmental science and the holistic approach . Lawrence Erlbaum Associates, Mahaw, pp 63–83
43. Kelso JAS (2005) Context, components and complexity. Invited paper, Plexus Institute Summit, Delray Beach, September 11
44. Kelso JAS (2007) The Haken–Kelso–Bunz Model. Scholarpedia (Computational Neuroscience/Dynamical Systems). http://www.scholarpedia.org/article/Haken-Kelso-Bunz_model
45. Kelso JAS, Engstrøm DA (2006) The complementary nature. MIT Press, Cambridge
46. Kelso JAS, Scholz JP, Schöner G (1986) Dynamics govern switching among patterns of coordination. Phys Lett A 134:8–12
47. Kelso JAS, Schöner G, Scholz JP, Haken H (1987) Phaselocked modes, phase transitions and component oscillators in coordinated biological motion. Physica Scripta 35:79–87
48. Kelso JAS, DelColle J, Schöner G (1990) Action-perception as a pattern formation process. In: Jeannerod M (ed) Attention and Performance XIII. Erlbaum, Hillsdale, pp 139–169
49. Kelso JAS, Bressler SL, de Guzman GC, Ding M, Fuchs A, Holroyd T (1991) Cooperative and critical phenomena in the human brain revealed by multiple SQUIDS. In: Duke D, Pritchards W (eds) Measuring chaos in the human brain. World Scientific, New Jersey, pp 97–112
50. Kelso JAS, Bressler SL, de Guzman GC, Ding M, Fuchs A, Holroyd T (1992) A phase-transition in human brain and behavior. Phys Lett A 169:134–144
51. Kelso JAS, Fuchs A, Lancaster R, Holroyd T, Cheyne D, Weinberg H (1998) Dynamic cortical activity in the human brain reveals motor equivalence. Nature 392:814–818
52. Kelso JAS, Fuchs A, Jirsa VK (1999) Traversing scales of brain and behavioral organization. I.-III. In: Uhl C (ed) Analysis of neurophysiological brain functioning. Springer, Heidelberg, pp 73–125
53. Kelso JAS, Fink PW, DeLaplain CR, Carson RG (2001) Haptic information stabilizes and destabilizes coordination dynamics. Proc R Soc Lond B Biol Sci 268:1207–1213
54. Kilner JM, Paulignan Y, Blakemore SJ (2003) An interference effect of observed biological movement on action. Curr Biol 13:522–525
55. Konner M (2004) The ties that bind - Attachment: the nature of the bonds between humans are becoming accessible to scientific investigation. Nature 429:705
56. Körding KP, Fukunaga I, Howard IS, Ingram JN, Wolpert DM (2004) A neuroeconomics approach to inferring utility functions in sensorimotor control. PLoS Biol 2:1652–1656
57. Kuramoto Y (1984) Chemical oscillations, waves, and turbulences. Springer, Berlin
58. Lagarde J, Kelso JAS (2006) Binding of movement, sound and touch: multimodal coordination dynamics. Exp Brain Res 173:673–688
59. Lagarde J, Kelso JAS, Peham C, Licka T (2005) Coordination dynamics of the horse-rider system. J Mot Behav 37:418–424
60. Lagarde J, de Guzman GC, Oullier O, Kelso JAS (2006) Interpersonal interactions during boxing: Data and model. J Sport Exerc Psychol 28: S108
61. Lee DN, Lishman JR (1975) Visual proprioceptive control of stance. J Hum Mov Stud 1:87–95
62. Marsh KL, Richardson MJ, Baron RM, Schmidt RC (2006) Contrasting approaches to perceiving and acting with others. Ecol Psychol 18:1–37
63. Mayville JM, Bressler SL, Fuchs A, Kelso JAS (1999) Spatiotemporal reorganization of electrical activity in the human brain associated with a timing transition in rhythmic auditory-motor coordination. Exp Brain Res 127:371–381
64. McGarva AR, Warner RM (2003) Attraction and social coordination: Mutual entrainment of vocal activity rhythms. J Psycholing Res 32:335–354
65. McNeill WH (1995) Keeping together in time. Harvard University Press, Cambridge
66. Mechsner F (2004) A perceptual-cognitive approach to bimanual coordination. In: Jirsa VK, Kelso JAS (eds) Coordination dynamics: Issues and Trends. Springer, Berlin, pp 177–195
67. Mechsner F, Kerzel D, Knoblich G, Prinz W (2001) Perceptual basis of bimanual coordination. Nature 414:69–73
68. Motter AE, Nishikawa T, Lai YC (2003) Large-scale structural organization of social networks. Phys Rev E 68:e036105-e036110
69. Mottet D, Guiard Y, Ferrand T, Bootsma RJ (2001) Two-handed performance of a rhythmical fitts task by individuals and dyads. J Exp Psychol Hum Percept Perform 27:1275–1286
70. Néda Z, Ravasz E, Vicsek T, Brechet Y, Barabasi AL (2000) Physics of the rhythmic applause. Phys Rev e 61:6987–6992
71. Néda Z, Ravasz E, Brechet Y, Vicsek T, Barabasi AL (2000) The sound of many hands clapping - Tumultuous applause can transform itself into waves of synchronized clapping. Nature 403:849–850
72. Nicolis G, Prigogine I (1977) Self-organization in non-equilibrium systems. Wiley, New York
73. Oullier O, Kelso JAS (2006) Neuroeconomics and the metastable brain. Trends Cogn Sci 10:353–354
74. Oullier O, Temprado JJ (2005) The structure of visual backgrounds modulates interpersonal coordination dynamics in a virtual environment. J Sport Exerc Psychol 27:S118-S119
75. Oullier O, Bardy BG, Stoffregen TA, Bootsma RJ (2002) Postural coordination in looking and tracking tasks. Hum Mov Sci 21:147–167
76. Oullier O, de Guzman GC, Jantzen KJ, Kelso JAS (2003) On context dependence of behavioral variability in inter-personal coordination. Int J Comput Sci Sport 2:126–128
77. Oullier O, de Guzman GC, Jantzen KJ, Lagarde JF, Kelso JAS (2004) Spontaneous interpersonal synchronization is modulated by the degree of visual coupling. J Sport Exerc Psychol 26:S11-S11
78. Oullier O, Bardy BG, Stoffregen TA, Bootsma RJ (2004) Task-specific stabilization of postural coordination during stance on a beam. Mot Control 8:174–187

79. Oullier O, Jantzen KJ, Steinberg FL, Kelso JAS (2005) Neural substrates of real and imagined sensorimotor coordination. Cereb Cortex 15:975–985
80. Oullier O, de Guzman GC, Jantzen KJ, Lagarde J, Kelso JAS (2005) Spontaneous interpersonal synchronization. In: Peham C, Schöllom WI, Verwey W (eds) European workshop on movement sciences: Mechanics-Physiology-Psychology. Sportverlag, Köln, pp 34–35
81. Oullier O, Kelso JAS, Kirman AP (2008) Neuroeconomics: A dynamical systems perspective. Rev Econ Pol 118:51–62
82. Oullier O, de Guzman GC, Jantzen KJ, Lagarde J, Kelso JAS (2008) Social coordination dynamics: Measuring human bonding. Soc Neurosci 3:178–192
83. Richardson MJ, Marsh KL, Schmidt RC (2005) Effects of visual and verbal interaction on unintentional interpersonal coordination. J Exp Psychol Hum Percept Perform 31:62–79
84. Richardson MJ, Marsh KL, Isenhower RW, Goodman JR, Schmidt RC (2007) Rocking together: dynamics of intentional and unintentional interpersonal coordination. Hum Mov Sci 26:867–891
85. Rizzolatti G, Craighero L (2004) The mirror-neuron system. Annu Rev Neurosci 27:169–192
86. Schmidt RC, O'Brien B (1997) Evaluating the dynamics of unintended interpersonal coordination. Ecol Psychol 9:189–206
87. Schmidt RC, Richardson MJ (2008) Dynamics of interpersonal coordination. In: Fuchs A, Jirsa VK (eds) Springer, Heidelberg, pp 281–308
88. Schmidt RC, Carello C, Turvey MT (1990) Phase transitions and critical fluctuations in the visual coordination of rhythmic movements between people. J Exp Psychol Hum Percept Perform 16:227–47
89. Schöner G, Kelso JAS (1988) Dynamic pattern generation in behavioral and neural systems. Science 239:1513–1520
90. Schurmann M, Hesse MD, Stephan KE, Saarela M, Zilles K, Hari R, Fink GR (2005) Yearning to yawn: the neural basis of contagious yawning. Neuroimage 24:1260–1264
91. Shockley K, Baker AA, Richardson MJ, Fowler CA (2007) Articulatory constraints on interpersonal postural coordination. J Exp Psychol Hum Percept Perform 33:201–208
92. Simon JR (1969) Reactions toward the source of stimulation. J Exp Psychol 81:174–176
93. Solnick SJ, Schweitzer ME (1999) The Influence of Physical Attractiveness and Gender on Ultimatum Game Decisions. Organ Behav Hum Decis Process 79:199–215
94. Sommerville JA, Decety J (2006) Weaving the fabric of social interaction: Articulating developmental psychology and cognitive neuroscience in the domain of motor cognition. Psychon Bull Rev 13:179–200
95. Stefan K, Cohen LG, Duque J, Mazzocchio R, Celnik P, Sawaki L, Ungerleider L, Classen J (2005) Formation of a motor memory by action observation. J Neurosci 25:9339–9346
96. Temprado JJ, Laurent M (2004) Attentional load associated with performing and stabilizing a between-persons coordination of rhythmic limb movements. Acta Psychol 115:1–16
97. Temprado JJ, Swinnen SP, Carson RG, Tourment A, Laurent M (2003) Interaction of directional, neuromuscular and egocentric constraints on the stability of preferred bimanual coordination patterns. Hum Mov Sci 22:339–363
98. Tognoli E, Magne C, de Guzman GC, Kelso JAS (2007) Brain rhythms underlying intentional social coordination. Society Neurosci Program 304
99. Tognoli E, Lagarde J, de Guzman GC, Kelso JA (2007) The phi complex as a neuromarker of human social coordination. Proc Nat Acad Sci USA 104:8190–8195
100. Turvey MT (2004) Impredicativity, dynamics and the perception-action divide. In: Jirsa VK, Kelso JAS (eds) Coordination Dynamics: Issues and Trends. Springer, Berlin
101. Vallacher RR, Nowak A (1997) The emergence of dynamical social psychology. Psychol Inquiry 8:73–99 (and commentaries therein)
102. van Ulzen NR, Lamoth CJ, Daffertshofer A, Semin GR, Beek PJ (2008) Characteristics of instructed and uninstructed interpersonal coordination while walking side-by-side. Neurosci Lett 432:88–93
103. Vinkovic D, Kirman A (2006) A physical analogue of the Schelling model. Proc Nat Acad Sci USA 103:19261–19265
104. Wallenstein GV, Kelso JAs, Bressler SL (1995) Phase transitions in spatiotemporal patterns of brain activity and behavior. Physica D 84:626–634
105. Wheeler ME, Peterson SE, Buckner RL (2004) Memory's echo: Vivid remembering reactivates sensory-specific cortex (vol 97, pg 11125, 2000). Proc Nat Acad Sci USA 101:5181
106. Winfree AT (1967) Biological rhythms and behavior of populations of coupled oscillators. J Theor Biol 16:15–42
107. Winfree AT (1980) The geometry of biological time. Springer, New York
108. Winfree AT (2002) On emerging coherence. Science 298: 2336–2337

Books and Reviews

Frith C, Wolpert DM (2004) The neuroscience of social interaction: Decoding, imitating and influencing the actions of others. Oxford University Press, Oxford

Fuchs A, Jirsa VK (eds) (2008) Coordination: Neural, behavioral and social dynamics. Springer, Berlin

Jantzen KJ, Kelso JAS (2007) Neural coordination dynamics of human sensorimotor behavior: A Review. In: Jirsa VK, McIntosh AR (eds) Handbook on brain connectivity. Springer, Berlin, pp 421–461

Jeannerod M (2006) Motor cognition: What actions tell to the self. Oxford University Press, Oxford

Jirsa VK, Kelso JAS (2004) Coordination dynamics: Issues and trends. Springer, Berlin

Kelso JAS (1995) Dynamic Patterns: The self-organization of brain and behavior. MIT Press, Cambridge

Kelso JAS, Engstrøm DA (2006) The complementary nature. MIT Press, Cambridge

Oullier O, Jantzen KJ (2008) Neural indices of behavioral instability in coordination dynamics. In: Fuchs A, Jirsa VK (eds) Springer, Berlin, pp 205–227

Pikovsky A, Rosenblum M, Kurths J (2001) Synchronization: A universal concept in nonlinear science. Cambridge University Press, Cambridge

Rizzolatti G, Craighero L (2004) The mirror-neuron system. Ann Rev Neurosci 27:169–192

Sommerville JA, Decety J (2006) Weaving the fabric of social interaction: Articulating developmental psychology and cognitive neuroscience in the domain of motor cognition. Psychon Bull Rev 13:179–200

Strogatz SH (2003) Sync: The emerging science of spontaneous order. Hyperion Press, New York

Social Network Analysis, Estimation and Sampling in

OVE FRANK
Statistics Department, Stockholm University,
Stockholm, Sweden

Article Outline

Glossary

Network model A network model describes how an observed network structure with its data could be generated from probabilistic assumptions.

Network population A network population is a finite or infinite collection of existing or conceivable networks. For instance, the population could consist of network instances of a process changing with time, or the population could consist of network realizations in different sets of units.

Observation scheme An observation scheme describes what variables are known or can be observed for units and relations in the population and in the sample.

Population network A population network is a population of units equipped with relational structure between them. Usually several attributes are attached both to units and to pairs of units, and the network can be described as a valued graph on the population of units with vertex and edge or arc variables defined in the graph.

Sampling design A sampling design describes how units are sampled from a population of units and what the selection probabilities are for all the possible sample sequences or sample subsets. The sampling procedure is usually controlled by the investigator. If the sampling is not controlled by the investigator but voluntary or "controlled by nature", the sample network can be considered as an outcome of a probabilistic network model, and there is no clear separation between design-based and model-based inference.

Sampling frame A sampling frame is a list of identifying labels of all possible sample units.

Social network A social network is a network consisting of units and relations of interest in the social sciences. The term is used both for population networks and for networks belonging to a network population.

Survey sampling Survey sampling of networks refer to the investigation of networks sampled from a network population while survey sampling in networks refer to the investigation of sub-networks sampled from a population network. Survey sampling of networks is usually model-based with an infinite network population. Survey sampling in networks is usually design-based with a finite population network.

Definition of the Subject

Sampling in network analysis can refer either to the sampling of networks from a population of networks or to the sampling of one or more sub-networks from a population network. Sampling designs and model approaches based on probabilistic methods allow statistical tools to be developed for the analysis of data obtained in surveys carried out in a network setting.

From about the 1960s the rapid development of computers and computer power changed the scene for statistics and applied mathematics in general. New possibilities of processing large amounts of data by using fast computer algorithms for calculations, sorting, and searching contributed to an accelerated interest in computer science and discrete mathematics. Simulation techniques and other computer intensive numerical methods revolutionized the arsenal of tools available in statistics and other parts of applied mathematics. Graph theory was established not only as an expanding part of discrete mathematics but also became applied as a tool for investigating relational structures in the social sciences. Friendship, co-operation, dominance, support and other relations between individuals were studied as sociograms or as more general networks in which both the individuals and the relations were attributed with qualitative and quantitative variables like gender and age of the individuals and type and strength of the relations. Other network applications include transmission of infectious diseases between people, work scheduling and other planning, routing systems for transportation of goods, traffic flows in urban areas, information flows in communication channels, administrative and organizational structure, capital transfer and surveillance systems. Characteristic of such network applications is the inherent complexity of co-variation or correlation that could exist within and between variables defined for different elements of the network.

Introduction

Special sampling designs using network information for successive link tracing or sample adjustment have been described in the literature under various names such as multiplicity sampling, snowball sampling, and Respondent driven sampling. Such designs have been used for investigations of hard-to-reach populations of various kinds, like drug addicts, homeless people and individuals at risk for hiv-infection, tuberculosis, or hepatitis. Such designs have also been proved useful in investigations of rare properties of individuals in populations that could be sampled from available frames but would require very large samples with conventional sampling designs.

This presentation focuses on sampling designs in finite population networks and network models developed for survey sampling of social networks. Estimation methods are described that can be applied under different assumptions about what population data are known and what sample data can be observed.

The next three sections give background material of various kinds. A brief historical background is given that describes the development of a survey sampling theory for networks. Basic terminology and some general notation needed is then specified and serves the purpose of unifying and simplifying some of the concepts and arguments used later. A common and characteristic feature of network surveys is that they allow non-orthodox sampling methods. Hence it is important to specify when such methods are justified. Some fundamentals about sampling designs are given together with an overview of some standard designs.

After this background follows five sections on network sampling designs. Sections "Snowball Sampling Designs"–"Estimation Based on Network Walk Samples" treat snowball sampling designs and the important sub-class of network walk designs. For various special cases and modifications of basic snowball and walk samples, standard estimation methods are discussed. Section "Methods for Estimating Hidden Populations" is devoted entirely to the estimation of sizes and properties of hidden populations and rare parts of large populations.

The discussion of network designs is followed by a presentation of various probabilistic network models used especially in the social sciences. Some of these models are appropriate in combination with designs or as tools for deriving so-called model-assisted estimators. The main interest in these network models stems from their use when an observed network is considered as an instant of a network process or as a sample network from a super-population of networks.

Next there are two sections on sampling designs and models for bipartite networks. Bipartite networks have units of two kinds which generally require very different treatment so that it is not appropriate to consider them as special cases of networks. Section "Bipartite Network Sampling Designs" describes sampling designs and estimators in bipartite networks. A combination of sampling designs and random modeling in bipartite networks is illustrated in Sect. "A Bipartite Network Model for Crime Participation" which treats a specific criminological application.

Finally, Sect. "Future Directions" comments briefly on possible future directions for the development of network surveys in the social sciences.

Background

Some early attempts to incorporate graphs in a statistical investigation are the studies on sociograms [49], on snowball sampling [33], and on sampling in graphs [3]. Of particular interest is a paper by Stephan [60] pointing to the need for network methods in sample survey theory. He suggests the term nexus sampling for surveys that consider data from units as well as from relations between units. Sampling in populations with graph structure is also discussed in the articles [6,14], and a comprehensive monograph on statistical inference in graphs is the thesis [16].

An independent discussion of some first steps in network sampling is given in [34,42]. Various problems in survey sampling and estimation in networks were discussed in a series of papers during the 1970s and 1980s: [7,17,18,19,20,21,22,23,24,25,26,27,48]. These papers use essentially various network sampling designs and no elaborate probabilistic model assumptions.

The theory of random graphs developed mainly from two models introduced by Erdös and Renyi in 1959 and 1960 [12,13], the uniform graph with n vertices and N randomly selected edges among the unordered pairs of vertices, and the Bernoulli graph with n vertices and independent insertions of edges with a common probability p for all the unordered pairs of vertices. These random graphs have challenged many mathematicians, and there is now an extensive literature on combinatorial and asymptotic results for random graphs. A brilliant comprehensive reference on general graph theory is [11]. Monographs concentrating on random graphs are [4,37,46].

Statistical inference based on probabilistic network models without explicitly addressing the sampling design problem has been treated for some variants of Bernoulli graphs, but generally such models have too little structure to be of interest in a statistical context with network data. A network model with a more elaborate and flexible parametrization that make it applicable in the social sciences is the random digraph model of Holland and Leinhardt introduced in [36]. A comprehensive monograph with many references to the social network literature is [67]. The introduction of Markov graphs in [30] and the log-linear modeling of random networks in [47,57,68] have had a major impact on the development of statistical methods for social network analysis. The edited volumes [5,8] are good sources for the methodological development. Various aspects of network modeling and statistical inference from network samples are the topics of several theses in recent years. Well known to me are the following important contributions to the methodological development: [10,35,38,40,42,55,59,62].

Basic Terminology and Notation

Survey sampling concepts of basic importance are populations of units, sample selection, sampling designs, observation schemes, and variables of interest defined on the units. Two standard reference works for survey sampling in finite populations are [54,63]. In order to make network sampling possible, the population has to be equipped with a relational structure between its units. Usually it is given as a binary variable defined on the ordered pairs of units. A population with a binary relation can be represented as a graph in which the vertices represent the units and the edges (arcs) indicate the unordered (ordered) pairs of units that are related. In a network setting vertices and edges (arcs) are sometimes called nodes and links.

The population from which samples are drawn can be different from the target population of the survey. It can also be changing with time so that particular precautions need to be taken. Disregarding such complications we assume that there is a finite population of N units labeled by integers $1, \dots, N$. The population is denoted $U = \{1, \dots, N\}$. Samples are drawn from U either as ordered sequences of units with or without repetitions allowed or as subsets of U. In the first case, a sample sequence of m units drawn is denoted $u = (u_1, \dots, u_m)$, and in the second case a sample subset is denoted S. Both u and S are random entities when the sample is drawn according to a probability design. The probability design of u is given by a probability distribution over the set U^m of all ordered m-sequences of units from U, and the probability design of S is given by a probability distribution over the class of all subsets of U. If the size n of S is fixed, all subsets of other sizes have zero probability.

The observation scheme specifies which variables are known or observable. A variable defined on the units (a vertex variable) is given by a function x from U to some range space R. Formally,

$$x = \{(i, x_i) \colon i \in U\} \,,$$

is a set of pairs (i, x_i) assigning a value x_i in R to each unit i in U. If there is a specified ordering of the units, x could also be represented as a sequence of values ordered in the same way, e. g. an ordered sequence $(x_1, \dots, x_N)$ corresponding to the order $1, \dots, N$ of the units. The first representation is more convenient when restrictions to different subsets of U are considered. For any subset A of U we denote by

$$x(A) = \{(i, x_i) \colon i \in A\} \,,$$

the restriction of the variable x to units in A. In particular, $x(U) = x$ and $x(\emptyset) = \emptyset$. A binary variable x with $R = \{0, 1\}$ and, more generally, a K-category variable x with $R = \{0, \dots, K-1\}$ partitions the population U into K disjoint subsets

$$U_j = \{i \in U \colon x_i = j\} \quad \text{for} \quad j \in R \,.$$

A variable defined on the ordered pairs of units (an arc variable) is given by a function y from U^2 to some range space R. Formally,

$$y = \{(i, j, y_{ij}) \colon i \in U, j \in U\} \,,$$

is a set of triplets (i, j, y_{ij}) assigning a value y_{ij} in R to each ordered pair (i, j) in U^2. If there is a specified ordering of the units in U, y could also be represented as a matrix array (y_{ij}) with N rows and N columns ordered according to the unit-ordering. For any two subsets A and B of U we denote by

$$y(A, B) = \{(i, j, y_{ij}) \colon i \in A, j \in B\} \,,$$

the restriction of the variable y to ordered pairs of units (i, j) with $i \in A$ and $j \in B$. The set $y(A, B)$ is simply referred to as the set of values of y from A to B. In particular, $y(U, U) = y$ and $y(A, \emptyset) = y(\emptyset, B) = \emptyset$ for any subsets A and B of U. A binary variable y with $R = \{0, 1\}$ and, more generally, a K-category variable y with $R = \{0, \dots, K-1\}$ partitions U^2 into K disjoint subsets

$$\{(i, j) \in U^2 \colon y_{ij} = r\} \quad \text{for} \quad r \in R \,.$$

Sampling Designs

In order to gain general information about population units with access only to the information from a sample of units, it is evident that the sample should represent the population in some specified way. Even if the population is partitioned into subpopulations according to several attributes of the units, and the sample is selected so that some units from each subpopulation are included in the sample, one cannot be sure that the sample allows inference to be drawn with confidence about collective properties of all the population units.

Survey sampling theory is based on information from samples that are selected according to probabilistic designs. A sampling design that gives all population units a positive probability of being included in the sample is required to achieve a sample that is representative of the population. The essential property of a sampling design that makes it appropriate for population inference is that its inclusion probabilities should be known or estimable for all units. This allows sample data to be weighted so that they accurately represent population data. A probabilistic sampling design not only makes inference possible; it also makes it possible to specify how inference uncertainty can be judged and quantified. During the early days of survey sampling, the benefits of probabilistic sampling designs were not always properly understood and had to be emphasized. Today this is so well known that it may seem dubious if someone tries to use non-probability samples.

Therefore, it is important to clarify the legitimacy of network sampling designs that involve arbitrarily or conveniently selected samples. The essential principle is still that inclusion probabilities should be known or estimable. Some network sampling designs are based on an initial probability sample followed by sequential samples that may be systematically or probabilistically selected. As a consequence, it might be hard to estimate inclusion probabilities for the final network sample, but there are no doubts about its legitimacy.

When the network sample is generated from an initial sample that is not a probability sample, it is important that the sequential samples generated after the initial sample are selected according to specified probabilistic rules. Under some fairly general assumptions to be specified, the inclusion probabilities for the final network sample can be proved to converge towards limits that are independent of the way the initial sample was selected. Thus, the lack of knowledge about the initial conditions is compensated for by letting the network sample consist of several waves of units sequentially selected according to a known or estimable probability design. This wave design determines the limiting inclusion probabilities of the network sampling design, and the accuracy of these probabilities increases with the number of waves. As we will see later, it is possible to modify the probabilistic wave design so that the limiting inclusion probabilities of the network sample become uniform or equal to any other desired probability distribution over the population of units.

It is convenient to specify a few standard sampling designs and their inclusion probabilities which are fundamental for some of the estimators discussed later. Consider a population U of N units and a sample subset S selected according to a uniform probability design over the n-subsets of U. This design is specified as

$$S \sim \text{Unif}(n, U) .$$

It is often called simple random sampling without replacement. Thus, the probability that $S = s$ is equal to

$$P(s) = n!/N(N-1)\dots(N-n+1) ,$$

for any n-subset s of U. The probability of inclusion in S of a fixed unit i from U is equal to the sum of the selection probabilities $P(s)$ over all s containing unit i. It is equal to n/N and denoted by

$$\pi_i = n/N ,$$

for any i in U. The probability of inclusion in S of two distinct units i and j from U equals

$$\pi_{ij} = n(n-1)/N(N-1) ,$$

for any distinct i and j in U.

Another subset design called Bernoulli sampling with parameter $p(0 < p < 1)$ is specified as

$$S \sim \text{Bern}(p, U) ,$$

and is defined by selection probabilities

$$P(s) = p^n(1-p)^{N-n} ,$$

for any subset s of U of size n for $n = 0, 1, \dots, N$. Its inclusion probabilities are

$$\pi_i = p ,$$

for any i in U, and

$$\pi_{ij} = p^2 ,$$

for any distinct i and j in U. An important property of Bernoulli sampling is that its inclusion indicators

$I_i = I(i \in S)$ are independent. The inclusion indicators are 1 or 0 according to whether unit i is included in S or in its complement $U - S$.

A generalization of Bernoulli sampling is obtained if the inclusion indicators are independent but not necessarily identically distributed Bernoulli variables. The sampling design with independent Bernoulli(p_i) distributed inclusion indicators for i in U is sometimes called Poisson sampling (perhaps referring to the fact that the random sample size is approximately Poisson distributed with parameter $\lambda = p_1 + \cdots + p_N$). We specify this design as a Bernoulli sampling with general parameter $p = \{(i, p_i)\colon i \in U\}$ satisfying $0 < p_i < 1$ for all i in U; in short

$$S \sim \text{Bern}(p_i\colon i \in U)\,.$$

Its selection probabilities are

$$P(s) = \prod_{i \in s} p_i \prod_{j \in U-s} (1 - p_j)\,,$$

for any subset s of U, and its inclusion probabilities are

$$\pi_i = p_i\,,$$

for any i in U, and

$$\pi_{ij} = p_i p_j\,,$$

for any distinct i and j in U.

A probability design for a sample sequence $u = (u_1, \ldots, u_m)$ of m units independently drawn according to a uniform distribution over U has selection probabilities

$$P(i_1, \ldots, i_m) = 1/N^m\,,$$

for any sequence $(i_1, \ldots, i_m)$ in U^m. This design specified by

$$u_k \sim \text{Unif}(U) \quad \text{independent for } k = 1, \ldots, m\,,$$

is often called simple random sampling with replacement. Its inclusion probabilities are given by

$$\pi_i = 1 - (1 - 1/N)^m\,,$$

for any i in U, and

$$\pi_{ij} = 1 - 2(1 - 1/N)^m + (1 - 2/N)^m\,,$$

for any distinct i and j in U.

A generalization of this design to m independent identically distributed (IID) draws according to a probability distribution with N arbitrary positive probabilities $p(i)$ summing to 1 for i in U has selection probabilities

$$P(i_1, \ldots, i_m) = p(i_1) \ldots p(i_m)\,,$$

for any sequence $(i_1, \ldots, i_m)$ in U^m. We specify this design by

$$u_k \sim \text{IID}(p(i)\colon i \in U) \quad \text{for } k = 1, \ldots, m\,.$$

Its inclusion probabilities are

$$\pi_i = 1 - [1 - p(i)]^m\,,$$

for any i in U, and

$$\pi_{ij} = 1 - [1 - p(i)]^m - [1 - p(j)]^m + [1 - p(i) - p(j)]^m\,,$$

for any distinct i and j in U. A further generalization would allow the sequence of independent draws to be made according to different distributions.

For any probability design of a sample sequence $u = (u_1, \ldots, u_m)$ we define the inclusion count or multiplicity m_i of unit i as the number of draws equal to that unit. The multiplicities sum to m, and a positive multiplicity of a unit indicates its inclusion in the sample sequence. The set of distinct units in a sample sequence u is denoted by $s(u)$. This is a random subset of U of size $n(u)$ and expected size

$$\text{E} \;\; n(u) = \sum_{i \in U} \pi_i\,,$$

and variance

$$\text{Var} \;\; n(u) = \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j)\,.$$

Any sample sequence $u = (u_1, \ldots, u_m)$ with $u_k \sim$ IID $(p(i)\colon i \in U)$ for $k = 1, \ldots, m$ has multiplicities $(m_1, \ldots, m_N)$ that are multinomially distributed with parameters m and $(p(1), \ldots, p(N))$. The number of distinct units $n(u)$ in u is a sum of independent Bernoulli(π_i)-variables, which is asymptotically Poisson($\sum_i \pi_i$). With dependent draws it is generally difficult to derive the simultaneous probability distribution of the multiplicities or the distribution of the number of distinct units in u.

Snowball Sampling Designs

Let the population $U = \{1, \ldots, N\}$ have a graph structure given by a binary variable y defined on the pairs of population units. Let the sets of units adjacent after and before unit i be

$$A_i = \{j \in U\colon y_{ij} = 1\} \quad \text{and} \quad B_i = \{j \in U\colon y_{ji} = 1\}\,,$$

for $i \in U$. For any subset s of U, we denote by $A(s)$ the union of the A_i for $i \in s$. A snowball sampling design starts with an initial subset S_0 of U. This subset can either be a convenience sample or the outcome of a probability sampling design with a known or unknown probability distribution P_0 over the class of all subsets of U. In the latter case there are selection probabilities $P_0(s) \geq 0$ summing to 1 for all subsets s of U. In the former case S_0 is equal to a fixed set s_0, and it is convenient to put $P_0(s_0) = 1$. Note that in either case we don't even know that all units in the population have positive probabilities of being included in the initial sample.

The snowball sample is obtained by successively extending the initial sample S_0 with units selected from the sets of units that are adjacent after previously joined units. More precisely, a first wave W_1 is selected according to a probability distribution over the class of subsets of

$$(U - S_0) \cap A(S_0)\ .$$

The union of the initial sample S_0 and the disjoint first wave W_1 is equal to the one-wave snowball S_1. Generally, the kth wave W_k is selected according to a probability distribution over the class of subsets of

$$(U - S_{k-1}) \cap A(W_{k-1})\ ,$$

for $k = 1, 2, \ldots$ with $W_0 = S_0$. The k-wave snowball S_k is the union of the $(k-1)$-wave snowball S_{k-1} and the kth wave W_k. The waves are disjoint and

$$W_k = S_k - S_{k-1} \quad \text{for} \quad k = 1, 2, \ldots$$

The k-wave snowball S_k is called saturated if the kth wave W_k is the last wave that is non-empty, so that the sequence of snowballs has reached an absorbing state

$$S_k = S_{k+1} = \cdots$$

According to the selections of the waves, it follows that the k-wave snowball S_k evolves as a stochastic process depending on the $(k-1)$-wave snowball S_{k-1} and the kth wave which depends on S_{k-1} and $W_{k-1} = S_{k-1} - S_{k-2}$. Thus, the snowball process S_k has a two-step memory: The probability distribution of S_k conditional on $S_0, \ldots, S_{k-1}$ depends on S_{k-1} and S_{k-2} only, for $k = 2, 3, \ldots$ Another equivalent way to express this is to say that the stochastic process with states consisting of the current snowball and its last wave (S_k, W_k) is a Markov chain.

The transition probabilities of this Markov chain are

$$\begin{aligned} P(S_{k+1} = s_{k+1}, W_{k+1} = w_{k+1} | S_k = s_k, W_k = w_k) \\ = P(W_{k+1} = w_{k+1} | S_k = s_k, W_k = w_k) \\ = P(w_{k+1} | s_k, w_k)\ , \end{aligned}$$

for s_{k+1} equal to the union of s_k and w_{k+1}. The Markov chain is assumed to be time-homogeneous so that the transition probabilities do not depend on the stage parameter k. The transition probabilities do depend on graph parameters

$$y(w_k, U - s_k) = \{(i, j, y_{ij}) \colon i \in w_k, j \in U - s_k\}\ ,$$

governing possible wave selections and possibly also on other parameters governing properties of the wave selections. Assume that saturation occurs at stage k or later. Then the sequence of snowball samples $(S_0, S_1, \ldots, S_k)$ has a probability distribution given by

$$\begin{aligned} P(S_0 = s_0, S_1 = s_1, \ldots, S_k = s_k) = P(s_0, s_1, \ldots, s_k) \\ = P_0(s_0) P(w_1 | s_0, w_0) P(w_2 | s_1, w_1) \\ \cdots P(w_k | s_{k-1}, w_{k-1})\ , \end{aligned}$$

for any strictly increasing subset sequence $(s_0, s_1, \ldots, s_k)$ with $s_j - s_{j-1} = w_j$ for $j = 1, 2, \ldots, k$ and $s_0 = w_0$.

The probability distribution of the k-wave snowball S_k is obtained by summing the probabilities $P(s_0, s_1, \ldots, s_k)$ over all strictly increasing subset sequences $(s_0, s_1, \ldots, s_{k-1})$ with s_{k-1} strictly included in s_k. The initial sample and each new wave need to contain at least one unit, so that s_k contains at least $k + 1$ units. If s_k contains more units, the sum is extended over all possible distributions of the non-initial units among the k waves.

In order to be able to perform such calculations, we need to specify the common sampling design of the waves. At stage k, the next wave W_{k+1} has a probability distribution over the class of subsets of that part of U that is not included in the current snowball S_k but is adjacent after its last wave W_k. Now the selection of a subset for W_{k+1} can be considered as obtained by the recruitment of new units made by units in W_k. Each unit i in W_k can recruit only from A_i, the set of units adjacent after unit i. Assume that the recruited set R_i is a Bernoulli sample with inclusion probabilities p_{ij} for j in A_i. The recruited sets R_i are assumed to be selected independently for all i in U. The union of the sets R_i recruited by the units i in W_k is the set $R(W_k)$. This set is a subset of $A(W_k)$, and the part of it that is not included in S_k is the new wave W_{k+1}. Thus

$$W_{k+1} = (U - S_k) \cap R(W_k)\ ,$$

where $R(W_k)$ as a union of independent Bernoulli samples is itself a Bernoulli sample conditional on W_k. The probability that unit j is included in $R(W_k)$ equals the probability that j is included in at least one of the R_i for i in W_k. This probability is denoted

$$p(W_k, j) = 1 - \prod (1 - p_{ij})\ ,$$

where the product is over i in W_k and j is a unit in $A(W_k)$. It follows that conditional on the current state of the Markov chain (S_k, W_k), the next wave W_{k+1} is a Bernoulli sample with inclusion probabilities

$$p(W_k, j) \quad \text{for} \;\; j \;\; \text{in} \;\; (U - S_k) \cap A(W_k) \, .$$

The transition probabilities of the Markov chain (S_k, W_k) can now be specified as

$$P(w_{k+1}|s_k, w_k) = \left[\prod p(w_k, j)\right]\left[\prod q(w_k, j)\right] \, ,$$

where the first product is over j in w_{k+1} and the second product is over j in $U - s_{k+1}$ with probabilities denoted $q(w_k, j) = 1 - p(w_k, j)$. The basic parameters here are the probabilities p_{ij} that unit i would recruit unit j. Only units j in A_i can be recruited, and only units i included in the initial sample or any of the waves will have an opportunity to recruit. We can interpret the parameters p_{ij} as probabilities of independent recruitment arcs in the population graph. When y is considered as specifying a fixed population graph, p_{ij} is zero whenever y_{ij} is zero. If all units in A_i should have the same probability of being recruited, we could put $p_{ij} = \alpha_i y_{ij}$ where α_i is a probability governing the recruitment or co-operation activity of unit i. If recruitment also depends on availability, status or attraction of the recruited unit, we could put $p_{ij} = \alpha_i \beta_j y_{ij}$. The parameters α_i and β_i for units i in U could be related to a categorical variable x defined on U. For a K-category variable x the parameters $\alpha_i = \alpha(x_i)$ and $\beta_i = \beta(x_i)$ have K possible values each, and the parametrization involves only $2K$ parameters.

Recruitment or selection of adjacent members of the population can be considered as a sampling performed by the investigator or by the respondent. The units of the population can be people that could be interviewed by the investigator. The units can also be geographical regions or other entities for which the adjacent units can either be observed directly or determined by interviews with some unit representative. In such cases when the investigator can get access to information about all units that are adjacent after a sampled unit, then the recruitment of further units can be made by the investigator. There are situations when information about adjacent units might be sensitive or embarrassing to the respondent. If the respondent is unwilling to release information about all adjacent units, a possibility might be to ask the respondent just to select some of them at random. Such respondent driven sampling might in some situations give more reliable data and could provide a reasonable way to avoid some problems of integrity. For a more thorough discussion of this, see, for instance, [53]. There are techniques developed with distribution of recruitment cards and various benefits to participants in order to help the investigator to get new waves of the sample. Even if respondent i provides only R_i and not A_i, information could be collected about out-degrees or other statistics needed for the respondent driven probability design of the wave sampling.

The combination of design based and model based probabilities are standard in survey sampling and should be so also when network sampling is applied. Respondent driven sampling could be such a case when a random graph approach would be appropriate for the wave sampling. An illustration of combination of design-based and model-based approaches to network surveys is given in [66].

Estimation Based on Snowball Samples

Estimation of population data from snowball sample data requires inclusion probabilities of all population units. Any fixed unit j is included in a k-wave snowball sample if it is included in the initial sample or in any of its k following waves. Since the initial sample and the waves are disjoint, their inclusion events are mutually exclusive, and the probability that the snowball includes j equals the sum of the probabilities that the initial sample or any of the waves include j:

$$P(j \in S_k) = P(j \in S_0) + P(j \in W_1) + \cdots + P(j \in W_k) \, .$$

Obviously this inclusion probability increases strictly with the number of waves until the snowball is saturated. If the initial sample is a convenience sample, its inclusion probability is 1 or 0 depending on whether j belongs to it or not. According to the results derived above for the conditional distribution of the next wave W_{k+1} given the current state (S_k, W_k), the conditional probability that a unit j not in S_k should be included in W_{k+1} is given by

$$p(W_k, j) = 1 - \prod (1 - p_{ij}) \, ,$$

where the product extends over units i in W_k. Therefore

$$P(j \in W_{k+1}) = \mathrm{E} \;\; [1 - I(j \in S_k)] p(W_k, j) \, ,$$

where the expectation is over outcomes of (S_k, W_k). The probability of inclusion of unit j in S_{k+1} is obtainable as the sum of the probabilities of inclusion in the waves, but usually it is easier to first consider the conditional probability of inclusion of j in S_{k+1} given the current state (S_k, W_k). This probability is given by $p(S_k, j)$ if we define $p(S_k, j) = 1$ for $j \in S_k$. This is achieved by defining $p_{jj} = 1$ for $j \in U$. Hence, the unconditional inclusion probability is given by

$$P(j \in S_{k+1}) = \mathrm{E}\ \ p(S_k, j) = 1 - \mathrm{E}\ \ \prod_i [1 - I(j \in S_k) p_{ij}] ,$$

where the expectation is over outcomes of S_k. Even if the initial sample and all the recruitment samples are independent Bernoulli samples, it doesn't follow that S_k is a Bernoulli sample, and we cannot carry out the expectation factor by factor to obtain a recurrence relation between single unit inclusion probabilities of the current and the next snowball. In fact, it can be shown that for distinct *i* and *j* in *U*

$$\pi_{ij}(k) \geq \pi_i(k) \pi_j(k) ,$$

where $\pi_j(k) = P(j \in S_k)$ and $\pi_{ij}(k) = P(i \in S_k \ \ \& \ \ j \in S_k)$ for $k = 0, 1, \ldots$ Equality holds true if and only if there are no units of positive probability of inclusion in S_0, which are *k* steps adjacent before both *i* and *j*. An approximation that deserves to be further investigated for snowballs with few waves is

$$\pi_j(k+1) = 1 - \prod_i [1 - \pi_i(k) p_{ij}] ,$$

or, equivalently, using logarithms

$$\log[1 - \pi_j(k+1)] = \sum_i \log[1 - \pi_i(k) p_{ij}] .$$

When the inclusion probabilities are not too close to 1, we can obtain a linear approximation to the recurrence relation by series expansions of the logarithmic functions:

$$\pi_j(k+1) = \sum_i \pi_i(k) p_{ij} .$$

Since the inclusion probabilities are fundamental for many estimators, it would be of interest to know if there are sampling designs for which this recurrence relation is an acceptable approximation.

Consider a one-wave snowball with Bernoulli samples as described above. There are several possibilities for data collection that need to be distinguished. Assume that the value of a vertex variable *x* and the degree are observed for sample units. If the units are identifiable and individual recruitments are observed, data consist of (i, x_i, y_i) for $i \in S_0$ and (j, x_j, y_j) for $j \in R_i$ and $i \in S_0$. Units have multiplicities 0 or 1 in S_0, but their multiplicities in S_1 might be larger and that applies also to units in S_0 if they are recruited in the first wave. If only collective recruitments are observed, data consist of (i, x_i, y_i) for $i \in S_0$ and (j, x_j, y_j) for $j \in R(S_0)$. In this case, multiplicities in S_1 can be 0, 1, or 2, and 2 applies only to units in S_0. If only collective recruitments not included in S_0 are observed, data consist of (i, x_i, y_i) for $i \in S_0$ and (j, x_j, y_j) for $j \in W_1$. Now multiplicities are never larger than 1. If units are not separated between S_0 and W_1, data consist of (i, x_i, y_i) for $i \in S_1$. Also in this case multiplicities are never larger than 1. If the units are not identifiable and individual recruitments are observed, data consist of (x_i, y_i) for $i \in S_0$ and (x_j, y_j) for $j \in R_i$ and $i \in S_0$ with no possibility to observe multiplicities and no possibility to reduce data to $R(S_0)$, W_1, or S_1. If only collective recruitments are observed, we cannot separate S_0 and W_1 and we don't know S_1 or any multiplicities. If only collective recruitments of units not included in S_0 are observed, we can separate S_0 and W_1 and we know S_1 and we know that no multiplicities are larger than 1. Without identities and without separation of S_0 and W_1, data consist of (x_j, y_j) for $j \in S_1$. Depending on the observation scheme available, different estimators can be used.

To illustrate various estimators, it is convenient to use snowball sample indicators $S_{kj} = I(j \in S_k)$ for $j \in U$ and $k = 0, 1, \ldots$ and recruitment indicators $z_{ij} = I(j \in R_i)$ for $i \in U$ and $j \in U$. According to our assumptions S_{0j} are Bernoulli(p_j) for $j \in U$, and z_{ij} are Bernoulli(p_{ij}) for $i \in U$ and $j \in U$, and all these variables are independent. Here $0 \leq p_j \leq 1$ for $j \in U$, and by allowing the values 0 and 1 initial convenience samples are possible. Recruitments are restricted by the presence of edges in the acquaintance graph, which implies that $0 \leq p_{ij} \leq y_{ij}$ for distinct *i* and *j* in *U*. For convenience we define $p_{jj} = 1$ for $j \in U$. Note that even if the acquaintance graph is undirected with $y_{ij} = y_{ji}$, the recruitment graph given by arc indicators z_{ij} is directed.

Consider the total $T = \sum x_i$ of a vertex variable *x* and the total $D = \sum y_i$ of degrees in the acquaintance graph. Let us specify $p_j = y_j / D$ and $p_{ij} = y_{ij} / y_i$ for $i \neq j$. Assuming *D* known and using only data from S_0, we have an unbiased estimator of *T* given by

$$T' = \sum_i (x_i S_{0i} / p_i) = \sum_i (x_i S_{0i} D / y_i) .$$

Similarly the total population size *N* has an unbiased estimator

$$N' = \sum_i (S_{0i} D / y_i) .$$

Without assuming *D* known, we have an asymptotically unbiased estimator of the population mean of *x* according to

$$T'/N' = \sum_i (x_i S_{0i} / y_i) \Big/ \sum_i (S_{0i} / y_i) .$$

Thus, the average of x-values in S_0 is weighted by reciprocal degrees.

Using individual recruitment data, an unbiased estimator of T is given by

$$\begin{aligned} T'' &= \sum_j \left(x_j \sum_i S_{0i} z_{ij} \Big/ \sum_i p_i p_{ij} \right) \\ &= \sum_i \sum_j (x_j S_{0i} z_{ij} D / y_j) \,. \end{aligned}$$

Also here the degree total D cancels from the asymptotically unbiased estimator of the population mean of x. We get

$$T''/N'' = \sum_i \sum_j (x_j S_{0i} z_{ij}/y_j) \Big/ \sum_i \sum_j (S_{0i} z_{ij}/y_j) \,.$$

Thus, the average of x-values in S_1 is counted with multiplicities and weighted by reciprocal degrees.

Using collective recruitments and only data from S_1, an unbiased estimator of T is given by

$$T''' = \sum_j [x_j S_{1j}/\pi_j(1)] \,,$$

where

$$\begin{aligned} \pi_j(1) &= 1 - \prod_i (1 - p_i p_{ij}) \\ &= 1 - \prod_i (1 - y_{ij}/D) = 1 - (1 - 1/D)^{y_i} \,. \end{aligned}$$

This time D does not appear as a factor in T''' or in the corresponding N''', so it does not cancel in the population mean estimator T'''/N'''. Replacing x_j by y_j in the estimators T', T'', and T''', we obtain D', D'', and D''' that all depend on D. The first two wouldn't provide any possibility to iteratively determine a value of a D-estimator. For D''' an iterative method would start with an initial value of D in $\pi_j(1)$, find D''', use this D''' for D in $\pi_j(1)$, find a new D''', etc. Convergence properties of such algorithms and properties of estimators that can be obtained from them haven't been explored. We will comment on similar iterative methods in Sect. "Estimation Based on Network Walk Samples".

When data from snowballs with more than one wave are available, it might be comparatively easy to set up estimators based on a first few waves, but more complicated to find the required inclusion probabilities for snowballs of many waves. An important technique described in [64] is to use Rao-Blackwellization to improve estimators based on data from only a few waves to obtain estimators based on all data. Rao-Blackwellization of an estimator requires extensive numerical calculations of its possible values for all arrangements of available data that are consistent with the reduced data remaining when multiplicities of sample units and order between the sample units are ignored. The expected value of an estimator conditional on a sufficient statistic in the form of the reduced data is the improved estimator. Sufficiency in survey sampling from a finite population was treated in [1]. More recent references are [9,44].

Network Walk Sampling Designs

Assume that the population $U = \{1, \ldots, N\}$ has a graph structure given by a binary variable y defined on the set of ordered pairs of population units. Let A_i and B_i be the sets of units adjacent after and before unit i. The size of A_i is the out-degree of unit i, and the size of B_i is the in-degree of unit i. When the graph is undirected, $A_i = B_i$ and out- and in-degrees are equal and called degrees.

A sampling design based on a random walk in a graph selects the sample as a sequence of units $u_1, u_2, \ldots$ according to a time-homogeneous Markov chain with transition probabilities

$$P(u_{k+1} = j | u_k = i) = P_{ij} \,,$$

that are independent of the stage parameter k for j in A_i and $k = 1, 2, \ldots$ If the graph is directed and strongly connected so that every unit has a directed path to every other unit, and if not all cycle lengths are multiples of a common factor larger than one, then the Markov chain will be irreducible and aperiodic. If the graph is undirected, connected, and contains at least one cycle of length one or three, then the Markov chain will be irreducible and aperiodic. A time-homogeneous, irreducible, and aperiodic Markov chain has a unique limiting distribution

$$\lim P(u_k = j | u_1 = i) = p_j > 0 \,,$$

independent of the initial unit i as k tends to infinity. The limiting distribution is equal to the stationary distribution that can be obtained by solving the equation system

$$p_j = \sum p_i P_{ij} \quad \text{for} \quad j = 1, \ldots, N \,,$$

subject to $\sum p_j = 1$. For instance, let the transition probability be $P_{ij} = y_{ij}/y_i$ for an undirected graph with degree y_i of unit i. This means that the transition from unit i is made with equal probabilities to each one of the y_i units that are adjacent to it. It follows that $p_i = y_i/(y_1 + \cdots + y_N)$. Hence, transition from any unit to one of its adjacent units according to a uniform probability distribution implies that the limiting probabilities of the units are proportional to their degrees.

If the limiting probabilities p_i for the units i in U cannot be determined from known or observable data, ergodic theory proves that they can be estimated by the relative frequencies in a long sample sequence obtained according to any time-homogeneous, irreducible, and aperiodic Markov chain. To be more specific, consider a sequence $u = (u_1, \dots, u_m)$ of length m. The multiplicities of the units are m_i for i in U. The relative multiplicity sequence $(m_1/m, \dots, m_N/m)$ converges towards the limiting distribution $(p_1, \dots, p_N)$ when m tends to infinity.

Let $n(u)$ be the size of the set $s(u)$ of distinct units in u. The probability π_i that unit i is included in $s(u)$ equals the probability that $m_i > 0$. The multiplicity sequence is a reduction of the information in the sample sequence u that ignores the order of selection. The set $s(u)$ is a further reduction that ignores both the order of selection and the multiplicities of the units selected.

The graph employed for random walk sampling determines the zero entries among the transition probabilities $P = (P_{ij})$. These network-induced transition probabilities determine the limiting stationary distribution $p = (p_1, \dots, p_N)$. If the initial unit u_1 is selected according to the stationary distribution p, this distribution applies to the marginal distribution of any position u_k of the walk. Another initial distribution implies that the stationary distribution is approached asymptotically on the walk for positions u_k as k increases. The limiting stationary distribution is important for appropriate weighting of data collected by network walk samples. Therefore, it is of special interest that it is possible to achieve any pre-assigned limiting distribution by adjusting the transition probabilities P induced by the network.

Assume that it is desirable to obtain a limiting distribution $q = (q_1, \dots, q_N)$ that is different from the stationary distribution p derived from the transition probabilities P. Modified new transition probabilities $Q = (Q_{ij})$ satisfying the reversibility condition $q_i Q_{ij} = q_j Q_{ji}$ would lead to the stationary distribution q, which is also the limiting distribution when transitions are made according to the new Q. If we define Q by

$$Q_{ij} = P_{ij} \min(1, q_j P_{ji} / q_i P_{ij}) \,,$$

for $i \neq j$ and

$$Q_{ii} = 1 - \sum_{i \neq j} Q_{ij} \,,$$

it follows that

$$q_i Q_{ij} = \min(q_i P_{ij}, q_j P_{ji}) = q_j Q_{ji} \,,$$

i. e. the reversibility condition is satisfied. To implement transitions according to Q, we can tentatively apply P and then decide to accept or reject a suggested transition from i to j with probability Q_{ij}/P_{ij} and $1 - Q_{ij}/P_{ij}$, respectively. More specifically, if the walk at stage k is in position $u_k = i$, and the next transition according to P would lead to unit j, the investigator might have to collect information from j to be able to calculate P_{ji} and compare it with P_{ij}. If $q_j P_{ji} \geq q_i P_{ij}$, then the transition to j has a probability valid also according to Q, and the walk is allowed to jump to $u_{k+1} = j$. Otherwise, the transition to j has not a valid probability according to Q, and, unless a transition is accepted, the walk has to stay at $u_k = i$ for another trial. Equivalently we could put $u_{k+1} = i$ and let the stage parameter count the number of trials instead of the number of jumps to other positions. Note that the probability that the modified walk stays for another trial equals the sum of the probabilities that the walk according to P jumps to a unit to be rejected, which is

$$\sum_{i \neq j} P_{ij}(1 - Q_{ij}/P_{ij}) = 1 - \sum_{i \neq j} Q_{ij} \,.$$

Consider the following illustrations for a directed network with transition probabilities $P_{ij} = y_{ij}/a_i$ for $j \in A_i$. Here the stationary probabilities can be difficult to find. Assume that we want a stationary distribution with probabilities q proportional to the out-degrees: $q_j = a_j/(a_1 + \cdots + a_N)$. The modified transition probabilities are

$$Q_{ij} = P_{ij} \min(1, q_j P_{ji} / q_i P_{ij}) = \min(y_{ij}, y_{ji})/a_i \,,$$

and only jumps to mutual contacts are accepted. We need to collect information about out-degrees and be able to verify mutuality. Assume instead that we want a stationary distribution proportional to in-degrees: $q_j = b_j/(b_1 + \cdots + b_N)$. Then

$$Q_{ij} = (y_{ij} y_{ji}/b_i) \min(b_i/a_i, b_j/a_j) \,.$$

Only jumps to mutual contacts are possible, and only those with $b_i/a_i \leq b_j/a_j$ have transition probabilities that are valid also according to Q. Thus, unit j can be rejected unless it is a mutual contact with at least the same ratio of out-degree to in-degree as unit i. In this case we need to collect information about out- and in-degrees and be able to verify mutuality. Finally, assume that a uniform limiting distribution $q_j = 1/N$ is wanted. Then

$$Q_{ij} = \min(P_{ij}, P_{ji}) = y_{ij} y_{ji} / \max(a_i, a_j) \,,$$

and j can be rejected unless it is a mutual contact that has at most the same out-degree as unit i. This case requires information about out-degrees and mutuality. Further illustrations are provided in [65]. A general presentation of methods for constructing transition kernels (probabilities or probability densities) with desired properties, including the Metropolis–Hastings algorithm and the Gibbs sampler, can be found in Sect. 3.6 of [39].

The random walk sampling discussed so far has been in undirected graphs that are connected or in directed graphs that are strongly connected. If the connectedness assumption is uncertain for the graph being employed, or if it is known that it is not valid, then the Markov chain is reducible, and the transitions cannot reach beyond the connected component induced by the initial sample. In order not to confine the chain to a part of the population only, transitions to adjacent units could sometimes be replaced by transitions to other units.

A way to control the balance between staying on the current walk and starting on a new one, is to use a mixture distribution for the transitions. Let $P' = (P'_{ij})$ be a matrix of probabilities for transitions to adjacent units, and let $P'' = (P''_{ij})$ be a matrix of probabilities for transitions to arbitrary units in the population. A mixture distribution with transition probabilities given by

$$P = \alpha P' + (1 - \alpha)P'' ,$$

where $0 < \alpha < 1$, is said to have a damping factor α measuring the proportion of transitions that are selected according to P'. For instance, $P'_{ij} = y_{ij}/a_i$ for j in A_i and $P''_{ij} = 1/N$ for j in U represent a mixture between uniform transitions from i to A_i and uniform transitions from i to U. The walk stops when it reaches a unit with $a_i = 0$. A slight modification is to define P'_{ij} also when $a_i = 0$, so that the walk is never forced to stop. For α close to 1, the damping effect will be small, and a population with a disconnected graph will be poorly represented by walk sample data. For decreasing α, the damping effect increases, the influence of the graph decreases, and the importance of having an appropriate matrix P'' increases. It might be appropriate to assume, for instance, that global transitions are made with probabilities proportional to out-degrees or in-degrees of the units, which would reflect that activity or attraction of the units is likely to influence the selections. In this way, the damping effect needed to handle a disconnected graph does not necessarily diminish the influence of the graph.

When network walk sampling is applied to webpage transitions in [70], a model is used that in our notation can be given as $P'_{ij} = y_{ij}/a_i$ if $a_i > 0$ and $P'_{ij} = 1/N$ if $a_i = 0$, and $P''_{ij} = p''_j$. Here p''_j reflects the web surfer's preferences for various pages, and the damping factor α reflects tendencies to use page links for surfing rather than own preferences.

Estimation Based on Network Walk Samples

Assume that sample data consist of the values of a variable x defined on the population of units. When units in the sample sequence can be identified, sample data consists of m values of x that can be referred to the units. The values (i, x_i, m_i) are known for all units i in $s(u)$. When units in the sample sequence cannot be identified, sample data consist of m values of x that cannot be referred to the units. Neither the multiplicities nor the set of distinct units can be determined. Sample data are restricted to the frequency distribution of the x-values.

There is an intermediate situation that could appear when sample data consist of the values of a variable x defined on U. If units in the sample sequence cannot be identified, it could still be possible to successively mark them in some way and observe when the same unit is sampled again. Data then consist of a sequence of m values on x that cannot be referred to the units but that can be partitioned into n sub-sequences of constant values belonging to the same unknown unit. Notice that the same value might appear in different sub-sequences. The multiplicities are known but their unit affiliations are not, and the number of distinct units is known but not their identities. The values (x_i, m_i) are known for n unknown distinct units i in U. If x is a K-category variable, the frequency distribution of these values (x_i, m_i) can be displayed in a K by m array representing available sample data when identities of units can not be observed but similarity between units is observed. Let $(n_1, \dots, n_m)$ be the marginal frequency distribution of the multiplicities. Then

$$n_1 + n_2 + \cdots + n_m = n ,$$

and

$$n_1 + 2n_2 + \cdots + mn_m = m .$$

Knowledge of the frequencies $(n_1, \dots, n_m)$ represents a reduction of the information in the multiplicity sequence that ignores the identities of the units. The marginal frequency distribution of the K-category variable x is the only information left when neither similarities nor identities between units can be observed.

Consider a network walk sample sequence $u = (u_1, \dots, u_m)$ obtained by a time-homogeneous, irreducible, and aperiodic Markov chain with transition probabilities $P_{ij} = y_{ij}/a_i$ for $j \in A_i$ and initial probabilities equal to the stationary probabilities $p_j = y_j/(y_1 + \cdots + y_N)$ for $j \in U$. The following reasoning applies asymptotically for large sample sequences even if the initial unit is not selected according to the stationary distribution. Data consist of the values of a vertex variable x and the degrees for each unit in the sample sequence u, but the units need not be identified. The population mean of x is $\mu = T/N$ where $T = x_1 + \cdots + x_N$. The mean of x in the sample sequence is

$$\sum_{i\in u}(x_i/m) = \sum_{i\in U}(m_i x_i/m)\ ,$$

where m_i is the unknown number of units equal to i in u. Since

$$\mathrm{E}\ m_i = mp_i = my_i/(y_1 + \cdots + y_N)\ ,$$

the sample mean is biased as an estimator of the population mean μ unless all degrees are equal. Now

$$T' = \sum_{i\in u}(x_i/mp_i) = \sum_{i\in U}(m_i x_i/mp_i)\ ,$$

has expected value T, and

$$N' = \sum_{i\in u}(1/mp_i) = \sum_{i\in U}(m_i/mp_i)\ ,$$

has expected value N. Moreover,

$$\mu' = T'/N' = \sum_{i\in u}(x_i/y_i)\Big/\sum_{i\in u}(1/y_i)\ ,$$

has asymptotically expected value μ for large values on m. The quantities T' and N' are not available from data since they contain the unknown total $D = y_1 + \cdots + y_N$ of all degrees. However, this total cancels in the ratio $\mu' = T'/N'$ so μ' is an asymptotically unbiased estimator of μ. We note that the estimator μ' is a weighted mean of the values of x in the sample sequence u with weights that are inversely proportional to the degrees.

The results obtained can be applied with $x_i = y_i$ to get an estimator of the average degree D/N in the population. Its estimator is

$$D'/N' = m\Big/\sum_{i\in u}(1/y_i)\ ,$$

which is the harmonic mean of the degrees in the sample sequence. The results can also be applied with $x_i = I(i \in U_1)$, i. e. with x as an indicator variable for a specific subset U_1 of U of unknown size N_1. Then $T = N_1$ and the relative size of U_1 has an asymptotically unbiased estimator

$$N_1'/N' = \sum_{i\in u}(x_i/y_i)\Big/\sum_{i\in u}(1/y_i)\ ,$$

which is the ratio between the sum of inverted degrees of units in the sample sequence that belong to U_1 and the sum of inverted degrees of all units in the sample sequence. The average degree in U_1 is given by

$$D_1/N_1 = \left[\sum_{i\in U} x_i y_i\right]\Big/\left[\sum_{i\in U} x_i\right]\ ,$$

where $x_i = I(i \in U_1)$, and it can be estimated by

$$D_1'/N_1' = \left[\sum_{i\in u} x_i\right]\Big/\left[\sum_{i\in u}(x_i/y_i)\right]\ ,$$

which is the harmonic mean of the degrees of the units in the sample sequence that belong to U_1. We note that the corresponding arithmetic mean is larger than or equal to the harmonic mean and therefore over-estimates the (arithmetic) population mean. This is because we oversample units with large degrees. The bias is reduced by taking the harmonic sample mean.

If the units in the sample sequence are identifiable, it is possible to observe the multiplicities and the set of distinct sample units $s(u)$. Should the inclusion probabilities $\pi_i = P(i \in s(u))$ be known, this would allow the population total T to be estimated by the unbiased estimator

$$T'' = \sum_{i\in s(u)} (x_i/\pi_i)\ .$$

Similarly, N would have an unbiased estimator

$$N'' = \sum_{i\in s(u)} (1/\pi_i)\ ,$$

and μ an asymptotically unbiased estimator

$$\mu'' = \left[\sum_{i\in s(u)} (x_i/\pi_i)\right]\Big/\left[\sum_{i\in s(u)} (1/\pi_i)\right]\ .$$

The inclusion probabilities $\pi_i = P(i \in s(u))$ are approximately $1 - \exp(-mp_i)$ and according to our assumptions about the Markov chain, the stationary probabilities are $p_j = y_j/D$ where $D = \sum y_j$ is the degree total. Now the average degree D/N was shown above to be estimated by the harmonic mean of the degrees in the sample sequence. Altogether, this allows us to set up the estimator N'' as a function of N, and by replacing N by N'' we might have a possibility to solve iteratively for N'' in the equations

$$N'' = \sum_{i\in s(u)} [1 - \exp 1\big/(-my_i/D')]\ ,$$

$$D' = mN''\Big/\sum_{i\in u}(1/y_i)\ .$$

Possibilities of finding estimators iteratively, when the inclusion probabilities are functions of some of the parameters we want to estimate, deserve further study.

Methods for Estimating Hidden Populations

Hidden populations usually refer to populations of human individuals sharing a stigmatizing, illegal, risky, or embarrassing behavior that make them hard to find. There are no lists or frames to sample them from. Investigation of such populations might be of interest to society and necessary for determining appropriate resources to be allocated for social support and other types of social action. Standard survey sampling methods that require population frames cannot be used to study homeless people, drug users, sexual workers, individuals exposed to criminal or medical risks, etc. Various alternative approaches are discussed in [61].

Convenience samples of such hidden populations might be obtained by social field workers or by the police. The information gained from such sources could indicate the need for further study and data collection to obtain reliable inference about the hidden population. Members of the hidden population that are approached by social field workers might not be unwilling to co-operate and give away information about other members of the hidden population. Several practical methods described in the literature use various kinds of payments and other benefits to individuals in the hidden population that help the investigator to reach other members of the population for interview. When respondents are willing to reveal which other members they know of, the investigator can select the waves of a snowball sample. Respondents might be more willing to co-operate if they are asked to select at random one or two of the other members they know of. Thus, respondent driven random walk or small wave snowball sampling can be applied to the population network.

In cases when members of hidden populations are not willing to reveal themselves, it might still be possible to get information about them by applying network sampling. Assume for instance that the hidden population of interest is part of a larger population that can be sampled conventionally. A respondent sampled from the larger population might without revealing names or identities be able and willing to give some information about how many members of the hidden subpopulation she knows about. Such local network information about the hidden subpopulation can sometimes be used to infer information about the entire hidden subpopulation. As an illustrative example, consider the following setup considered in [41].

The hidden population U_1 consists of an unknown number N_1 of individuals. In order to estimate N_1, a larger population U of N individuals is defined so that it comprises U_1 as a subset; U_1 is embedded in U in such a way that another subset U_2 of U can be assumed to be equally common in U_1 and in U. The subset U_2 should consist of individuals having some easily identified characteristic. Neither the size N of U nor the size N_2 of U_2 need to be known. Select a sample S from U with positive inclusion probabilities π_i for all i in U. Each respondent is asked whether she belongs to U_2 and whether she has some acquaintances in U_2. She is not asked whether she belongs to U_1 but only whether she has some acquaintances in U_1. Let x_1 and x_2 be indicator variables of the sets U_1 and U_2, i. e. $x_{jk} = I(j \in U_k)$ for $j = 1, \dots, N$ and $k = 1, 2$. Let further $y = (y_{ij})$ be the symmetric adjacency matrix of the acquaintance graph with degree y_i equal to the unknown number of acquaintances of individual i in U. This number is usually quite large and the respondent need not be asked to estimate it but rather to concentrate on giving reliable estimates of her numbers of acquaintances that belong to U_1 and U_2, say

$$a_{ik} = \sum_{j \in U} x_{jk} y_{ij} ,$$

for $i = 1, \dots, N$ and $k = 1, 2$. If the independence assumptions are met for x_1, x_2, and y, then it could be expected under a randomization model that proportionality is valid according to

$$a_{ik}/y_i = N_k/N \quad \text{for } k = 1, 2 \text{ so that } a_{i1}/a_{i2} = N_1/N_2 .$$

The sum of the numbers a_{ik} for $i \in U$ equals the degree sum in U_k, i. e.

$$\sum_{i \in U} a_{ik} = \sum_{j \in U} x_{jk} y_j = D_k .$$

The standard Horvitz–Thompson estimator of D_k is

$$D'_k = \sum_{i \in S} (a_{ik}/\pi_i) ,$$

for $k = 1, 2$. Similarly Horvitz–Thompson estimators of N and N_2 are given by

$$N' = \sum_{i \in S} (1/\pi_i) \quad \text{and} \quad N'_2 = \sum_{i \in S} (x_{i2}/\pi_i) .$$

From the proportionality assumption we get that the average degrees are equal in U, U_1, and U_2:

$$D/N = \sum_{i \in U} y_i/N = D_k/N_k ,$$

for $k = 1, 2$. Hence, N_1 can be estimated by

$$N'_1 = D'_1 N'_2 / D'_2 .$$

By summing $a_{i1}/a_{i2} = N_1/N_2$ over i in U_2 and over i in U we get

$$\sum_{i \in U} (a_{i1}x_{i2}/a_{i2}) = N_1 \quad \text{and} \quad \sum_{i \in U} (a_{i1}/a_{i2}) = NN_1/N_2\,,$$

which leads to two alternative estimators of N_1, namely

$$N_1'' = \sum_{i \in S} (a_{i1}x_{i2}/a_{i2}\pi_i)$$

$$\text{and} \quad N_1''' = (N_2'/N') \sum_{i \in S} (a_{i1}/a_{i2}\pi_i)\,.$$

Here the sums are understood to be over sample units $i \in S$ with acquaintances in U_2. Should $a_{i2} = 0$ we expect $y_i = 0$ and $a_{i1} = 0$. If the three estimators N_1', N_1'', and N_1''' turn out to give very different estimates, the assumptions underlying the calculations are doubtful and another approach is needed.

This example illustrates that properties of hidden groups that are hard to access can still be estimated if the hidden group is embedded in a convenient network. While snowball sampling and random walk sampling use the network to successively extend an initial sample, the hidden group example uses randomization assumptions about stable expected average network degrees as a possible bridge between different subgroup sizes. If the network assumptions needed for the randomization procedure are not valid, then a natural alternative approach would be to combine the probability design of the sampling with a probability model of the network.

An entirely different kind of hidden population are animal populations. In wildlife surveys the technique of catch-recapture sampling is an attempt to handle the lack of population frames. Bird counting, for instance, relies on marking the birds caught in a first sample by identifying rings before they are released again. After some time a new sample is taken, and the number of marked birds in the first sample is divided by the proportion of marked birds in the second sample to get an estimator of the size of the bird population. This idea can be elaborated on by specifying more carefully various assumptions about how the population changes with time and how the risk of being caught depends on different environmental factors that can be controlled for. Catch-recapture techniques have also been discussed as a tool in criminology surveys to estimate the number of offenders of specific types of crimes in different regions and in different periods of time. Such surveys cannot be carried out without a thorough model specification as a complement to the sampling design. In Sect. "Bipartite Network Sampling Designs" an illustration is given of utilizing network models and network sampling methods in a criminological setting.

Area sampling is another technique that can be used for counting animal populations and other populations distributed over geographical areas. A population region is partitioned into a number of area sites. Neighboring sites can be considered as adjacent vertices in a graph. The size of the area population is a vertex variable, and the total population size is a vertex variable total that can be estimated by a snowball sample of area sites with counts of their area populations.

Probabilistic Network Models

Survey sampling in finite populations distinguishes between design-based and model-based inference. Design-based inference uses probabilistic sampling designs but considers population data as fixed, while model-based inference also makes probabilistic assumptions about population data. Essential properties of population data might be summarized in a few parameters in an appropriate model and thereby allow a convenient and simplified description of the population. The graph employed in a network sampling design can be exposed to various unknown influences and uncontrolled effects that make it natural to try to capture its essential features in a random graph model. Random graph models can also be used to describe the outcome when a graph is sampled from a population of graphs, or when an observed graph is generated by some complex process, for instance, being an instant realization of a network changing with time. Random graphs useful for statistical data analysis need to be sufficiently rich in parameters so that meaningful fits can be made. However, too many degrees of freedom in the model make it vulnerable to accidental fits with no explanatory power.

A simple Bernoulli graph on $U = \{1, \ldots, N\}$ with a symmetric adjacency matrix $y = (y_{ij})$ has $N(N-1)/2$ independent edge indicators y_{ij} that are Bernoulli(p)-distributed for $1 \leq i < j \leq N$. The only parameter p represents edge density. A block model variant of the Bernoulli graph has a K-category vertex variable x and assumes y_{ij} to be independent and Bernoulli distributed with a parameter $p(x_i, x_j) = p(x_j, x_i)$ for $1 \leq i < j \leq N$. Now there are $(K+1)K/2$ parameters representing edge densities within and between different blocks of vertices. Depending on whether block affiliation is known, can be observed, or is not observable, the possibilities to estimate the parameters are very different. Statistical inference for stochastic block models is considered in many papers. An extensive account with many references can be found in [67].

For a directed graph on U with adjacency matrix y there are $N(N-1)/2$ dyads (y_{ij}, y_{ji}) for $1 \leq i < j \leq N$. The simple model with IID dyads has four parameters

$p_{00}, p_{01}, p_{10}, p_{11}$ that are probabilities $p_{kl} = P(y_{ij} = k, y_{ji} = l)$, and the probabilities sum to 1 leaving 3 degrees of freedom. More generally, with independent dyads and probabilities $P_{ijkl} = P(y_{ij} = k, y_{ji} = l)$, there are $2N(N-1)$ parameters subject to $N(N-1)/2$ restrictions $\sum_k \sum_l P_{ijkl} = 1$, which leaves $3N(N-1)/2$ degrees of freedom. Assuming a block structure with a K-category vertex variable x and $P_{ijkl} = p_{kl}(x_i, x_j) = p_{lk}(x_j, x_i)$ there are 2 degrees of freedom within each category of size larger than one and 3 degrees of freedom between any pair of distinct categories, in total $2K + 3K(K-1)/2 = K(3K+1)/2$ degrees of freedom. In particular, $x_i = i$ for $i = 1, \ldots, N$ implies that each category consists of only one vertex, and there are 3 degrees of freedom for each pair of vertices, i. e. $3N(N-1)/2$ degrees of freedom in total in accordance with the previous result. Thus, a categorization of 20 vertices based on two binary vertex variables reduces the degrees of freedom from 570 to 26, and a simple dichotomy reduces the degrees of freedom to 7, so block models represent a substantial simplification of dyad independence models. A practical approach might be to start with a rather large number of tentative vertex variables and apply cluster analysis to their dyad distributions in order to find how many of these distributions need to be distinguished. This technique is applied in [31,32].

The block model approach can be considered even if the category affiliation is uncertain and described by independent vertex variables x_i with a probability distribution $P(x_i = k) = p_k$ for $k = 1, \ldots, K$. Let N_k be the number of vertices in category k and N_{kl} the number of pairs of vertices in categories k and l, i. e. $N_{kl} = N_k N_l$ for $k \neq l$ and $N_{kk} = N_k(N_k - 1)/2$. The log-likelihood function when vertex variables and graph adjacencies are observed is given by

$$\log L = \sum_{k=1,\ldots,K} N_k \log p_k + \sum\sum_{1\leq k\leq l\leq K} \sum_{m=0,1} \sum_{n=0,1} N_{klmn} \log p_{mn}(k,l)\,,$$

where

$$N_{klmn} = \sum\sum_{1\leq i<j\leq N} I(x_i = k, x_j = l, y_{ij} = m, y_{ji} = n)\,.$$

This model has $K-1$ degrees of freedom in addition to the previous $K(3K+1)/2$, so in total $3(K+1)K/2 - 1$ degrees of freedom. This model is treated for instance in [45,58,69].

Consider now the model with independent dyads and probabilities $P_{ijkl} = P(y_{ij} = k, y_{ji} = l)$. A convenient way to re-parametrize this model is to introduce odds and an odds ratio. Let a_{ij} be the odds of y_{ij} when $y_{ji} = 0$, and let b_{ij} be the odds of y_{ij} when $y_{ji} = 1$:

$$a_{ij} = P_{ij10}/P_{ij00}, b_{ij} = P_{ij11}/P_{ij01}\,.$$

The ratio between these odds is

$$c_{ij} = b_{ij}/a_{ij} = P_{ij11}P_{ij00}/P_{ij01}P_{ij10}\,.$$

The corresponding odds of y_{ji} are

$$a_{ji} = P_{ji10}/P_{ji00} = P_{ij01}/P_{ij00}\,,$$
$$b_{ji} = P_{ji11}/P_{ji01} = P_{ij11}/P_{ij10}\,,$$

and the odds ratio is the same

$$c_{ji} = b_{ji}/a_{ji} = c_{ij}\,.$$

The three parameters a_{ij}, a_{ji}, c_{ij} are arbitrary non-negative numbers, and together with the normalizing constant

$$P_{ij00} = 1/(1 + a_{ij} + a_{ji} + a_{ij}a_{ji}c_{ij})\,,$$

we retain the other probabilities as

$$P_{ij10} = a_{ij}P_{ij00}\,, \quad P_{ij01} = a_{ji}P_{ij00}\,,$$
$$P_{ij11} = a_{ij}a_{ji}c_{ij}P_{ij00}\,.$$

Finally we put

$$\lambda_{ij} = -\log P_{ij00}\,, \quad \alpha_{ij} = -\log a_{ij}\,, \quad \gamma_{ij} = -\log c_{ij}\,,$$

where $\lambda_{ij} = \lambda_{ji} > 0$ and $\gamma_{ij} = \gamma_{ji}$. Using the three parameters $\alpha_{ij}, \alpha_{ji}, \gamma_{ij}$ and the normalizing constant λ_{ij} determined by them, it is possible to express the likelihood function L according to

$$-\log L = \sum\sum_{1\leqslant i<j\leq N} (\lambda_{ij} + \alpha_{ij}y_{ij} + \alpha_{ji}y_{ji} + \gamma_{ij}y_{ij}y_{ji})\,.$$

There are $3N(N-1)/2$ degrees of freedom represented by $\alpha_{ij}, \alpha_{ji}, \gamma_{ij}$ for $1 \leq i < j \leq N$. By introducing an out-degree effect α_i, an in-degree effect β_i, and a mutuality effect γ_i for each vertex, we assume that

$$\alpha_{ij} = \alpha_i + \beta_j\,, \gamma_{ij} = \gamma_i + \gamma_j\,.$$

Since $\alpha_{ij} = \alpha + \beta + \alpha_i - \alpha + \beta_j - \beta$ for arbitrary α and β, α_i and β_j are identifiable only up to a translation

$\alpha = -\beta$. If we introduce an arc density effect δ and put $\alpha_{ij} = \delta + \alpha_i + \beta_j$ we can impose the restrictions $\sum \alpha_i = 0$ and $\sum \beta_j = 0$. In this way we have reduced the degrees of freedom from $3N(N-1)/2$ to $3N - 1$, and the likelihood L is a log-linear function of the parameters given by

$$-\log L = \sum_{1 \le i < j \le N}\sum \lambda_{ij} + \delta \sum_{i \ne j}\sum y_{ij} + \sum_i \alpha_i \sum_j y_{ij} + \sum_j \beta_j \sum_i y_{ij} + \sum_i \gamma_i \sum_j y_{ij} y_{ji} ,$$

with out-degree, in-degree, and number of mutual arcs at each vertex as the sufficient statistics. A further reduction is obtained by assuming all mutuality effects equal, $\gamma_i = \gamma$, which leads to a model with $2N$ degrees of freedom. This is the original Holland–Leinhardt model for directed graphs introduced and investigated in [36]. Various extensions to valued graphs and block models are possible, and algorithms for model fitting and parameter estimation are described and illustrated in the reference book [67] and in several references given there.

Random graph models with probability distributions of exponential type have been suggested with more elaborate network statistics. Some examples of investigations of such models are [47,52,57,68].

Instead of specifying which statistics are likely to be essential for the application considered and use an exponential type model with these statistics as sufficient statistics, Frank and Strauss in [30] derived the sufficient statistics from basic assumptions about the dependence structure in the adjacency matrix. They specified a Markov dependence structure for the adjacency matrix y and proved that it implied a log-linear likelihood with certain sufficient statistics. The Markov assumption is that y_{ij} and y_{kl} are independent whenever i, j, k, l are all distinct, but dependence is possible between any arc indicators with some common vertex. The sufficient statistics are then given by star and triangle statistics. Adding a homogeneity assumption saying that isomorphic structures have the same probability, the sufficient statistics are given by star counts and triangle counts. Here stars and triangles should be understood in a broad sense so that they can consist of dyads of any type. The fundamental tool for this result in [30] is a theorem of Besag given in [2] together with a dependence graph specifying the dependencies among arc indicators in y. This dependence graph is in fact the so-called line-graph of the graph given by y. The maximal cliques of the dependence graph are the minimal sufficient statistics of y. Further exploration of these ideas can be found in the thesis [50] and the article [51].

Bipartite Network Sampling Designs

A bipartite network consists of vertices of two kinds and vertex and edge variables defined on them. If U is a set of N vertices of the first kind, and V is a set of M vertices of the second kind, different vertex variables could be defined on U and V. Edge variables are defined between U and V. As a simple example, take U as a set of households, V as a set of individuals, and define a binary edge variable y_{ij} that indicates whether individual j belongs to household i for $i \in U$ and $j \in V$. Each individual belongs to exactly one household. Assume that U is a target population, and that we are interested in some variable x defined on U. Only a sample S from V can be obtained. For each individual sampled, the value of x is reported that belongs to the individual's household. Thus

$$\sum_{i \in U} x_i y_{ij} ,$$

is reported by $j \in S$. If the sampling design has inclusion probabilities

$$\pi_j = P(j \in S) ,$$

the household total $T = \sum_{i \in U} x_i$ has a design unbiased estimator

$$T' = \sum_{j \in S}\sum_{i \in U} x_i y_{ij}/\pi_j = \sum_{i \in U} x_i \sum_{j \in S} y_{ij}/\pi_j .$$

If more than one individual from the same household are sampled, the estimator contains this household value x with multiplicity. If household identities are checked and no repeated household values should be used, another design unbiased estimator is available as

$$T'' = \sum_{i \in B(S)} x_i/P(i \in B(S)) ,$$

where $B(S)$ is the set of distinct households reported by individuals in S. The probability $P(i \in B(S))$ is equal to $P(S \cap A_i \neq \emptyset)$, where A_i is the set of individuals belonging to household i. This is the probability that none of those individuals are included in the sample S. If we introduce multiplicities m_i for the number of individuals in S belonging to household i, there is an alternative design unbiased estimator

$$T''' = \sum_{i \in U} x_i m_i/\mathrm{E}m_i ,$$

that requires the expected value of the multiplicity. Comparisons between these estimators and further results on bipartite network sampling can be found in Sirken's pioneering paper on multiplicity sampling [56].

The possibility to use bipartite networks in survey sampling is not confined to cases with sampling units that are different from the units in the target population. As an illustration of a more complex combination of sampling from both U and V, we consider in the next section a model approach to a simultaneous estimation of crime rates and offender activity that relies on a bipartite network model.

A Bipartite Network Model for Crime Participation

Consider crimes of a particular kind in a specified geographical region during a specified period of time, say assaults in Canada during year 2000. Let M denote the unknown total number of incidents and m the number known to the police. Thus, $M - m$ is an unknown so-called "dark figure" of incidents, and m is the number of identified incidents. The identified incidents involve r incidents reported to the police and $m - r$ incidents identified by other means. Let m_i be the number of identified incidents having i identified offenders, and let r_i be the number of reported incidents having i identified offenders for $i = 0, 1, \ldots$ Incidents with at least one identified offender are said to be cleared by the police. There are unidentified offenders for all the unidentified incidents, and there are unidentified co-offenders for the identified incidents. The unknown numbers of incidents M_{ij} with i actual offenders and j identified offenders for $j = 0, 1, \ldots, i$ and $i = 1, 2, \ldots$ involve many dark numbers in addition to $M - m$. The unknown numbers of incidents

$$M_i = \sum_{j=0,\ldots,i} M_{ij} \,,$$

with i actual offenders for $i = 1, 2, \ldots$ provide a partition of M that specifies the distribution of incidents according to size, i. e. their number of co-offenders.

The same offender can be involved in several offences. Let N denote the unknown total number of distinct individuals participating in the M offences. The known number n of distinct offenders identified in the cleared incidents imply that there is a dark number of $N - n$ non-identified offenders. The number of incidents in which an offender participates is called the activity of the offender. We need to distinguish between identified activity and actual activity. Let N_{ij} be the number of offenders with activity i and identified activity j for $j = 0, 1, \ldots, i$ and $i = 1, 2, \ldots$. The partition of N into the numbers

$$N_i = \sum_{j=0,\ldots,i} N_{ij} \,,$$

of offenders of activity i for $i = 1, 2, \ldots$ specifies the distribution of offenders according to activity. This distribution involves further dark numbers that cause its discrepancy from the known numbers n_j of identified offenders participating in j identified offences for $j = 1, 2, \ldots$

In order to estimate the distribution of incidents according to size and the distribution of offenders according to activity, a bipartite network model can be defined with the set U of M incidents as vertices of the first kind and the set V of N offenders as vertices of the second kind. Criminal participation can be specified by an edge from $i \in U$ to $j \in V$ if incident i involves offender j. Let A_i be the subset of V specifying the co-offenders of incident i, and let B_j be the subset of U specifying the re-offences of offender j.

Now we can consider the set of identified incidents as a sample S from U, and the set of identified co-offenders of identified incident $i \in S$ as a sample S_i from A_i. These samples have sampling designs depending on factors governing how incidents are reported to the police or identified by other means and how the co-offenders are identified. In a situation like this when sampling is not determined by the investigator it is natural to incorporate the sampling design in a probabilistic network model.

An approach taken in [28] specifies the incidents in U as generated by a homogeneous Poisson process with parameter λ representing the expected number of incidents per time unit. The sample S of identified incidents is a Bernoulli sample with inclusion probability α, which means that identified incidents occur according to a homogeneous Poisson process with parameter $\alpha\lambda$. The parameter α measures the proportion of identified incidents. At least for crimes with a victim, α can be estimated by comparing incidence reporting frequencies and incidence identifications by other means with incidence frequencies estimated from victimization studies.

The set V of offenders is assumed to consist of a fixed number N of individuals that have a positive probability of offending. Independently for each incident $i \in U$, a set A_i of co-offenders in that incident is selected from V according to some common probability sampling design with inclusion probabilities $\pi_j > 0$ for $j \in V$. The sampling design might involve unknown parameters reflecting what is known about structure and behavior in the population V. For each identified incident $i \in S$, a set S_i of identified co-offenders is modeled as an independent Bernoulli sample of A_i with parameter β governing the chance that the police can identify an offender. An optimistic and a pessimistic view of this chance distinguishes between whether or not the identification of one co-offender in an incident should imply the identification of all co-offenders in that incident.

With this model, it turns out to be possible to estimate simultaneously both the distribution of incidents according to size and the distribution of offenders according to activity. The basic results are the following. From the distribution of identified size among identified incidents, it is possible to estimate the parameter β and the distribution of true size among all incidents. Furthermore, if θ is the average identified activity among identified offenders and $\gamma = n/N$ is the unknown proportion of identified offenders, then identified activity has a Poisson$(\gamma\theta)$ distribution that is truncated at zero, and true activity has a Poisson$(\gamma\theta/\alpha\beta)$ distribution. This implies that

$$\gamma = 1 - \exp(-\theta\gamma) ,$$

from which γ can be solved, and N estimated by n/γ.

An interesting feature of this bipartite network model is that it makes explicit assumptions about both incident size and offender activity, and these assumptions lead to estimators of all the dark numbers related to unidentified incidents, unidentified offenders, and unidentified co-offenders in identified incidents. In particular, the estimator of the hidden population size N of offenders provides a special insight into how this estimator can be modified if not all parts of the distribution of identified activity are reliable. This also makes interesting connections with the literature that derives or suggests such modifications by using other approaches.

Future Directions

It is likely that the development of more sophisticated network models for specific applications will appear as improvements of Markov models and other network models belonging to the exponential family. Such development is also expected to contribute to more efficient algorithms for computer intensive estimation and goodness-of-fit testing of network models.

The developments of design-based inference methods that are appropriate for sampling of sub-networks from a finite population network are especially challenging for few-waves snowball sampling and for network walk sampling. The iterative estimation methods discussed in Sects. "Estimation Based on Snowball Samples" and "Estimation Based on Network Walk Samples" are examples of problems that deserve to be investigated further. More important is perhaps the need to develop efficient algorithms for Rao-Blackwellization of various initially obtained estimators and to gain more information about minimal sufficient statistics in various snowball sampling designs.

Bibliography

1. Basu D, Ghosh JK (1967) Sufficient statistics in sampling from a finite universe. Proc 36th Session Int Stat Inst, pp 850–859
2. Besag JE (1974) Spatial interaction and the statistical analysis of lattice systems. J Royal Stat Soc B 36:192–236
3. Bloemena AR (1964) Sampling from a graph. Math Centre Tracts, Amsterdam
4. Bollobas B (2001) Random graphs. Cambridge University Press, Cambridge
5. Brandes U, Erlebach T (eds) (2005) Network analysis. Springer, Berlin
6. Capobianco M (1970) Statistical inference in finite populations having structure. Trans New York Acad Sci 32:401–413
7. Capobianco M, Frank O (1982) Comparison of statistical graph-size estimators. J Stat Plan Inference 6:87–97
8. Carrington PJ, Scott J, Wasserman S (eds) (2005) Models and methods in social network analysis. Cambridge University Press, Cambridge
9. Cassel CM, Särndal CE, Wretman J (1993) Foundations of inference in survey sampling. Krieger Publ Comp, Malabar
10. Corander J (2000) On Bayesian graphical model determination. Dissertation, Stockholm University
11. Diestel R (2005) Graph theory. Springer, Berlin
12. Erdös P, Renyi A (1959) On random graphs I. Publ Math Debrecen 6:290–297
13. Erdös P, Renyi A (1960) On the evolution of random graphs. Publ Math Inst Hungar Acad Sci 5:17–61
14. Frank O (1969) Structure inference and stochastic graphs. Swedish Research Institute of National Defence, Stockholm, FOA-Reports 3(6):1–10
15. Frank O (1970) Sampling from overlapping subpopulations. Metrika 16:32–42
16. Frank O (1971) Statistical inference in graphs. Dissertation, Stockholm University
17. Frank O (1977) Estimation of graph totals. Scand J Stat 4:81–89
18. Frank O (1977) A note on Bernoulli sampling in graphs and Horvitz–Thompson estimation. Scand J Stat 4:178–180
19. Frank O (1977) Survey sampling in graphs. J Stat Plan Inference 1:235–264
20. Frank O (1978) Sampling and estimation in large social networks. Soc Netw 1:91–101
21. Frank O (1978) Estimation of the number of connected components in a graph by using a sampled subgraph. Scand J Stat 5:177–188
22. Frank O (1979) Estimating a graph from triad counts. J Stat Comput Simul 9:31–46
23. Frank O (1979) Estimation of population totals by use of snowball samples. In: Holland P, Leinhardt S (eds) Perspectives on Social Network Research. Academic Press, New York, pp 319–347
24. Frank O (1980) Sampling and inference in a population graph. Int Stat Rev 48:33–41
25. Frank O (1980) Estimation of the number of vertices of different degrees in a graph. J Stat Plan Inference 4:45–50
26. Frank O (1981) A survey of statistical methods for graph analysis. In: Leinhardt S (ed) Sociological methodology. Jossey-Bass, San Francisco, pp 110–155
27. Frank O (1987) Random sampling and social networks – a survey of various approaches. Math Inform Sci Hum 26(104):19–33

28. Frank O, Carrington PC (2007) Estimation of offending and co-offending using available data with model support. J Math Sociol 31:1–46
29. Frank O, Snijders T (1994) Estimating the size of hidden populations using snowball sampling. J Off Stat 10:53–67
30. Frank O, Strauss D (1986) Markov graphs. J Am Stat Assoc 81:832–842
31. Frank O, Komanska H, Widaman K (1985) Cluster analysis of dyad distributions in networks. J Classif 2:219–238
32. Frank O, Hallinan M, Nowicki K (1985) Clustering of dyad distributions as a tool in network modelling. J Math Sociol 11:47–64
33. Goodman LA (1961) Snowball sampling. Ann Math Stat 32:148–170
34. Granovetter M (1976) Network sampling: Some first steps. Am J Sociol 81:1287–1303
35. Hagberg J (2003) On degree variance in random graphs. Dissertation, Stockholm University
36. Holland PW, Leinhardt S (1981) An exponential family of probability distributions for directed graphs. J Am Stat Assoc 76:33–65
37. Janson S, Luczak T, Rucinski A (2000) Random graphs. Wiley, New York
38. Jansson I (1997) On statistical modeling of social networks. Dissertation, Stockholm University
39. Kaipio J, Somersalo E (2005) Statistical and computational inverse problems. Springer, New York
40. Karlberg M (1997) Triad count estimation and transitivity testing in graphs and digraphs. Dissertation, Stockholm University
41. Killworth PD, McCarty C, Bernard HR, Johnsen EC, Domini J, Shelley GA (2002) Two interpretations of reports of knowledge of subpopulation sizes. Soc Netw 25:141–160
42. Koskinen J (2004) Essays on Bayesian inference for social networks. Dissertation, Stockholm University
43. Morgan DL, Rytina S (1977) Comment on "Network sampling: Some first steps, by M Granovetter". Am J Sociol 83:722–727
44. Mukhopadhyay P (2001) Topics in survey sampling. Springer, New York
45. Nowicki K, Snijders TAB (2001) Estimation and prediction for stochastic blockstructures. J Am Stat Assoc 96:1077–1087
46. Palmer EM (1985) Graphical evolution. Wiley, New York
47. Pattison P, Wasserman S (1998) Logit models and logistic regressions for social networks. II: Multivariate relations. Br J Math Stat Psychol 52:169–193
48. Proctor CH (1979) Graph sampling compared to conventional sampling. In: Holland PW, Leinhardt S (eds) Perspective on social network research. Academic Press, New York, pp 301–318
49. Proctor CH, Loomis CP (1951) Analysis of sociometric data. In: Jahoda M, Deutsch M, Cook SW (eds) Research methods in social relations. Dreiden Press, New York, pp 561–586
50. Robins GL (1998) Personal attitudes in inter-personal contexts: Statistical models for individual characteristics and social relationships. Dissertation, University of Melbourne
51. Robins GL, Pattison P (2005) Interdependencies and social processes: Dependence graphs and generalized dependence structures. In: Carrington PJ, Scott J, Wasserman S (eds) Models and methods in social network analysis. Cambridge University Press, Cambridge, pp 192–214
52. Robins GL, Pattison P, Wasserman S (1999) Logit models and logistic regressions for social networks. III: Valued relations. Psychometrika 64:371–394
53. Salganik MJ, Heckathorn DD (2004) Sampling and estimation in hidden populations using respondent-driven sampling. Sociol Methodol 34:193–239
54. Särndal CE, Swensson B, Wretman J (1992) Model assisted survey sampling. Springer, New York
55. Schweinberger M (2007) Statistical methods for studying the evolution of networks and behaviour. Dissertation, University of Groningen
56. Sirken MG (1970) Household surveys with multiplicity. J Am Stat Assoc 63:257–266
57. Snijders TAB (2002) Markov chain Monte Carlo estimation of exponential random graph models. J Soc Struct 3(2)
58. Snijders TAB, Nowicki K (1997) Estimation and prediction of stochastic blockmodels for graphs with latent block structure. J Classif 14:75–100
59. Spreen M (1999) Sampling personal network structures: Statistical inference in Ego-graphs. Dissertation, University of Groningen
60. Stephan FF (1969) Three extensions of sample survey technique. In: Johnson NL, Smith H Jr (eds) New developments in survey sampling. Wiley, New York
61. Sudman S, Kalton G (1986) New developments in the sampling of special populations. Ann Rev Sociol 12:401–429
62. Tallberg C (2003) Bayesian and other statistical approaches for analyzing network block-structures. Dissertation, Stockholm University
63. Thompson ME (1997) Theory of sample surveys. Chapman & Hall, London
64. Thompson SK (2006) Adaptive web sampling. Biometrics 62(4):1224–1234
65. Thompson SK (2006) Targeted random walk designs. Surv Methodol 32:11–24
66. Thompson S, Frank O (2000) Model-Based Estimation with Link-Tracing Sampling Designs. Surv Methodol 26:87–98
67. Wasserman S, Faust K (1994) Social network analysis. Cambridge University Press, Cambridge
68. Wasserman S, Pattison P (1996) Logit models and logistic regressions for social networks. I: An introduction to Markov random graphs and p*. Psychometrika 60:401–426
69. Wellman B, Frank O, Espinoza V, Lundquist S, Wilson C (1991) Integrating individual, relational and structural analysis. Soc Netw 13:223–249
70. Wills RS (2006) Google's page rank: The math behind the search engine. Math Intell 28(4):6–11

Social Network Analysis, Graph Theoretical Approaches to

WOUTER DE NOOY
University of Amsterdam, Amsterdam,
The Netherlands

Article Outline

Glossary

Adjacent Two vertices are adjacent if they are connected by a line.

Arc An arc is a directed line. Formally, an arc is an ordered pair of vertices.

Attribute An attribute is a characteristic of a vertex measured independently of the network.

Bipartite network See: Two-mode network.

Clique A clique is a maximal complete subnetwork containing three vertices or more.

Complete A (sub)network is complete if it has maximum density: All possible lines occur.

Component A (weak) component is a maximal (weakly) connected subnetwork.

Degree The degree of a vertex is the number of lines incident with it.

Density Density of a simple network is the number of lines, expressed as a proportion of the maximum possible number of lines.

Digraph A digraph or directed graph is a graph containing one or more arcs.

Distance The distance from vertex u to vertex v is the length of the geodesic from u to v.

Edge An edge is an undirected line. Formally, an edge is an unordered pair of vertices.

Ego-network The ego-network of a vertex contains this vertex, its neighbors and all lines among the selected vertices.

Geodesic A geodesic is the shortest path between two vertices.

Graph A graph is a set of vertices and a set of lines between pairs of vertices.

Incident A line is defined by its two endpoints, which are the two vertices that are incident with the line.

Indegree The indegree of a vertex is the number of arcs it receives.

Line A line is a tie between two vertices in a network: Either an arc or an edge.

Loop A loop is a line that connects a vertex to itself.

Neighbor A vertex that is adjacent to another vertex is its neighbor.

Network A network consists of a graph and additional information on the vertices or the lines of the graph.

One-mode network In a one-mode network, each vertex can be related to each other vertex.

Outdegree The outdegree of a vertex is the number of arcs it sends.

Path A path is a semipath with the additional condition that none of its lines is an arc of which the end vertex is the arc's tail.

Reachable We say that a vertex is reachable from another vertex if there is a path from the latter to the former.

Semicycle A semicycle is a closed semipath ending at the vertex at which it starts.

Semipath A path is a closed sequence of lines such that the end vertex of one line is the starting vertex of the next line and no vertex in between the first and last vertex of the sequence occurs more than once.

Signed graph A signed graph is a graph in which each line carries either a positive or a negative sign.

Simple graph A simple undirected graph contains neither multiple edges nor loops. A simple directed graph does not contain multiple arcs.

Star-network A star-network is a network in which one vertex is connected to all other vertices but these vertices are not connected among themselves.

Strong component A strong component is a maximal subnetwork in which each pair of vertices is connected by a path.

Strongly connected A (sub)network is strongly connected if each pair of vertices is connected by a path.

Structural property A structural property is a characteristic (value) of a vertex that is a result of network analysis.

Triad A triad is a (sub)network consisting of three vertices.

Two-mode network In a two-mode network, vertices are divided into two sets and vertices can only be related to vertices in the other set.

Undirected graph An undirected graph does not contain arcs: All of its lines are edges.

Vertex (vertices) A vertex (singular of vertices) is the smallest unit in a network.

Weakly connected A (sub)network is weakly connected if each pair of vertices is connected by a semipath.

Definition of the Subject

Social network analysis (SNA) focuses on the structure of ties within a set of social entities or actors, e. g., persons, groups, organizations, and nations, or the products of human activity or cognition such as semantic concepts, web sites, and so on. In a graph theoretical approach, a social network is conceptualized as a graph, that is, a set of vertices (or nodes, units, points) representing social actors and a set of lines representing one or more social relations among them.

A network, however, is more than a graph because it contains additional information on the vertices and lines. Characteristics of the social actors, for instance a person's sex, age, or income, are represented by discrete or continuous attributes of the vertices in the network, and the intensity, frequency, valence, or type of social relation are represented by line weight or value, line sign, or line type. Formally (see pp. 94–95, 127–128 in [1]), a network **N** can be defined as $\mathbf{N} = (U, L, F_U, F_L)$ containing a graph $\mathbf{G} = (U, L)$, which is an ordered pair of a unit or vertex set U and a line set L, extended with a function F_U specifying a property of the units ($f\colon U \to X$) and a function F_L specifying a property of the lines ($f\colon L \to Y$). The set of lines L may be regarded as the union of a set of undirected edges E and a set of directed arcs $A (L = E \cup A)$. Each element e of E is an unordered pair of units u and v (vertices) from U, that is, $e(u\colon v)$, and each element a of A is an ordered pair of units u and v (vertices) from U, that is, $a(u\colon v)$.

The application of graph theory to social relations can be traced back to at least the 1940s (see pp. 69–72 in [2]) when the mathematician R. Duncan Luce and the engineer Albert Perry teamed up with the social psychologist Leon Festinger [3] and when the mathematician Frank Harary started his collaboration with Leon Festinger and afterwards with Dorwin Cartwright [4]. They extended pioneering work in SNA that had been done notably in sociometry [5,6] and anthropology [7,8,9]. In the 1960s, advances in graph theoretical approaches to SNA such as the contributions by Øystein Ore [10], Claude Flament [11], Frank Harary [12], and innovative applications such as Everett M. Rogers' work on the diffusion of innovations [13], prepared the ground for the rise of SNA in both the USA [14] and Europe [15] as a new set of methods or a new methodology [16] in the 1970s.

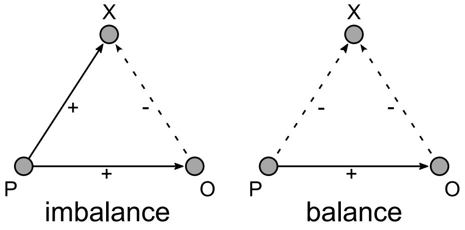
Introduction

The conceptualization of social systems as graphs and networks offered the opportunity for systematic investigation and theorizing of the structure of ties among social actors beyond the pair. Whereas classical sociology tended to make a quantum leap from the individual and the pair to the triple, group, or society [17], graph theory offered the tools to formally describe and visualize social structure consisting of three and more actors. This led to a new awareness of social structure as a system of ties that is both the product of human action and the context and condition for human action. Because this point of view is very relevant to the issue of complexity in social networks, it is briefly presented in the next paragraph.

The prevalent action theory in SNA conceptualizes collective behavior as the socially 'orchestrated' behavior of individuals or other actors. Actors adjust their behavior and attitudes, opinions, and beliefs to the behavior (etc.) of other members of the social system in which they participate. The system itself is not supposed to behave but it constrains actor behavior: It is the social context within which actors operate. As a network of ties, the system defines to whom an actor is exposed. The immediate contacts – the neighbors in graph theory – of an actor are usually most important to its behavior, but indirect contacts such as the neighbors' neighbors may be taken into account as well. In other words, an actor's local context or ego-network is likely to affect its behavior. At the same time, however, by ending ties or creating new ones, the individual changes both local network structure and overall network structure, that is, the system. Thus, individual action changes the local context for its neighbors' (neighbors' etc.) action. Complexity arises in the interplay between individual behavior and the system both as the overall structure of the network of social ties and as the local context for each actor within the system. To the actors, the change of network structure is not necessarily predictable, so the interplay between individual action and network structure may offer surprising results.

Let us turn to an example now, which is one of the earliest applications of graph theory to social networks. This example nicely illustrates both the transition from a focus on the tie within a single pair to the study of group structure in the social sciences and the interplay between local action and overall network structure. In 1946, the psychologist Fritz Heider formulated the theory of psychological balance [18], which stated among other things that a person (P) feels uncomfortable when he or she disagrees with his or her friend (O) on a particular topic (X). Person P is hypothesized to be stressed and to try to change this situation either by adapting its opinion on topic X, so it matches O's opinion, or by changing its opinion on O, regarding him or her no longer as a friend. Figure 1 represents both a situation of imbalance and a situation of balance as a signed digraph. The circles and arrows represent the vertices and arcs of the graph and the valence of the opinions or affections are shown both by the labels and the style of the arrows: Solid arcs show positive opinions or sentiments, dotted arcs show negative opinions or sentiments.

In 1956, the mathematician Frank Harary and the psychologist Dorwin Cartwright realized that psychological



Social Network Analysis, Graph Theoretical Approaches to, Figure 1
The principle of imbalance and balance conceptualized as signed digraphs

balance in this triad (three vertices and the lines among them) may be conceptualized as a specific pattern of arcs in a signed graph, viz., in a balanced triad, the $P - O - X$ semicycle (a closed semipath) always contains zero or an even number of negative arcs, whereas an imbalanced triad is characterized by a semicycle with an uneven number of negative arcs [4]. As a next step, replacing the topic by a third person and perceptions of liking or disliking (by the focal person P) by expressed liking or disliking as ties, they easily generalized this idea to a network of arbitrary size. They proved that a signed network is balanced if and only if all semicycles contain no or an even number of negative arcs.

In addition, they proved that a balanced network either contains one set of vertices with just positive arcs among them, or two sets of vertices with all positive arcs within the sets and all negative arcs between the sets, which is a polarized network. In 1967, this result was generalized to polarization among three or more groups by James A. Davis, who showed that a network can be partitioned into an arbitrary number of subsets of vertices such that all positive ties are within the subsets and all negative arcs are among the subsets if the network does not contain semicycles with exactly one negative line [19]. Figure 2 shows an example from Samuel F. Sampson's [20] study of a network of sentiments among novices in a monastery. It depicts the situation at the fourth measurement wave, which was highly polarized at that time. Vertex color indicates whether the novice had previously attended one particular seminary (black: Yes, white: No).

In this case, a mathematical relation was established between a behavioral hypothesis at the level of the individual, viz., adjusting ones affect relations such that ones situation is balanced, and overall network structure, viz., polarization. Individual behavior proved to have an unexpected outcome at the level of overall network structure and graph theoretical concepts, in this case semicycles, provided the link. In many cases, however, the link between individual behavior and overall network structure has not been established formally and is sometimes even hard to predict intuitively.

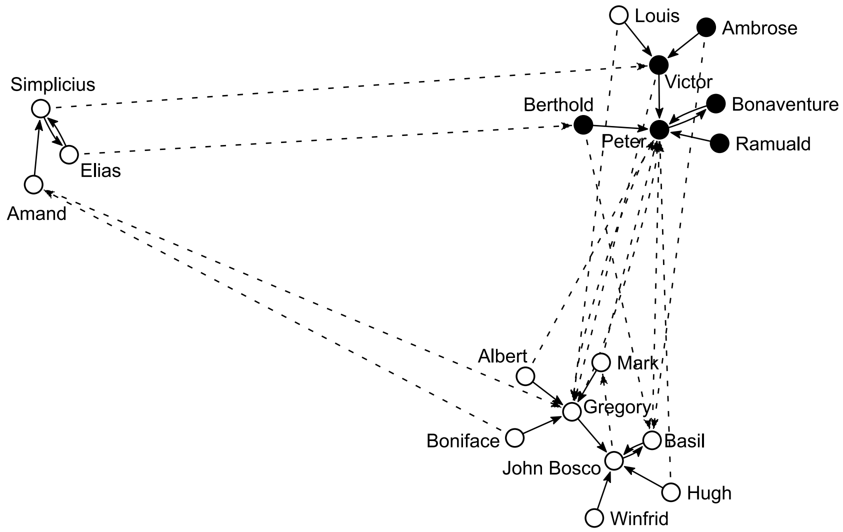
This entry of the encyclopedia aims to present an overview of graph theoretical approaches to SNA, highlighting the complex relations between individual action and overall network structure. It intends to explain why current developments focus on local structure rather than overall network structure to unravel the complexity of social networks. For each of the main topics in SNA (cohesion, brokerage, and prestige), behavioral hypotheses are presented stating why actors create, maintain or dissolve ties. The typical local structure of ties produced by this behavior is sketched in combination with graph theoretical indices for measuring it, and finally the expected consequences to the overall structure of the social network are discussed in combination with the graph theoretical measures developed for measuring them. Note that this approach more or less reverses the historical development of SNA, which focused on overall network structure first and gradually became more interested in the behavior of actors that created, maintained, or changed network structure.

Cohesion

One of the first intuitions in SNA concerns the tendency of human beings to form cohesive subgroups. This is a classical topic in the social sciences, see, for instance, George C. Homans' book *The Human Group* [21], and it was central to the sociometry tradition [22]. But where do cohesive subgroups come from and what do they do?

The first and most general behavioral hypothesis merely states that similar people tend to interact more easily and people who interact tend to become or perceive themselves as more similar provided that the interaction is characterized as positive, friendly, and so on. In SNA, this tendency is mainly known as homophily, a concept coined by Paul F. Lazarsfeld and Robert K. Merton [23,24], but it is known under other names in several scientific disciplines, e. g., the phenomenon of attribution [25] and affect control [26,27] in social-psychology, assortative or selective mixing in epidemiology and ecology ([28], p. 2 in [29]), and assortative mating in genetics with efforts at statistical modeling at least as early as 1985 [30].

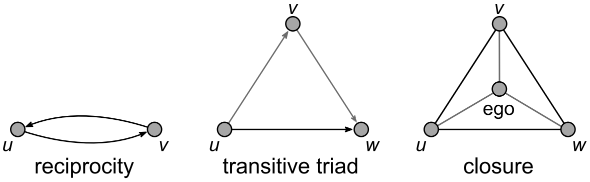
It is important to note that there are two sides to this behavioral hypothesis, a selection effect, that is, the impact of similarities on the ties that are created, sustained, or broken [31], and an influence effect [32], which hypothesizes that perceived or actual similarities such as the socially constructed identities or opinions [33] result from



Social Network Analysis, Graph Theoretical Approaches to, Figure 2
Almost perfect polarization in the network of sentiments among novices

the ties among actors. According to these hypotheses, people who are directly linked are or become more similar because of their interaction, so they become more likely to engage in ties and maintain ties among them. Thus, social groups form and persist as tightly linked sets of people that tend to share social and psychological characteristics, producing solidarity.

If we concentrate on the graph theoretical aspects of this behavioral hypothesis, that is, the structure of ties, and ignore the (dis)similarities among the actors for the moment, we find several characteristics of local structures that measure cohesive subgroup formation. At the level of a pair of actors, reciprocity of ties in directed networks signals subgroup formation: Both actors are hypothesized to choose each other when they are similar. At the level of the triple, transitivity results from tendencies toward cohesion. If actor u establishes a tie with actor v because they are similar, and actor v establishes a tie to actor w for the same reason, actors u and w are also similar, so actor u is expected to establish a tie with w as well, creating a so-called transitive triad (Fig. 3). Stated differently, the path from u via v to w is closed by an arc from u to w. In general, the closure of paths or semipaths both in directed and undirected networks signals cohesive subgroup formation at the local level. Closure within an ego-network may be calculated as the percentage of all possible ties among a vertex' neighbors that are present, which is one of the definitions of the clustering coefficient (see pp. 32–33 in [34]) but the concept of closure can be extended beyond a vertex' immediate neighbors, e. g., the number of semicycles of length 4 or larger in which a vertex is involved, e. g., balanced semicycles in signed networks.



Social Network Analysis, Graph Theoretical Approaches to, Figure 3
Reciprocity, transitivity, and closure in a directed ego-network

If we include vertex attributes in our measures of cohesive subgroup formation, we can calculate homophily quite simply as the probability or ratio of ties between vertices that share a particular characteristic to ties between vertices that do not. Extending this idea to the ego-network, the homogeneity of actors involved in an ego-network may be taken as a measure of tendencies toward homophily. For qualitative attributes of the actors, Blau's index of variability or heterogeneity [35] can be used ($1 - \sum p_i^2$) where p is the proportion of group members in a particular category and i is the number of different categories), which is conceptually related to the Herfindahl–Hirschman Index in economics measuring the extent of monopoly within an industry. It is interesting to note that Blau's theory hypothesizes that heterogeneity rather than homogeneity of actors within a group enhances the operation and efficiency of the group. If improving group efficiency is the aim of actors, we would have to use a behavioral hypothesis that is quite the opposite of the homophily hypothesis.

Tendencies toward cohesive subgroup formation at the level of actor behavior are most likely to produce sets of densely connected vertices in the overall network and, if we add attributes of the actors, the vertices within sets tend to have similar characteristics. In the most extreme case, the sets are disconnected, so the network consists of several weak components, that is, maximal weakly connected subnetworks, or they are connected only by ties with a negative social meaning, as in the example of polarization presented in the Introduction.

In the history of SNA, the concept of relatively densely connected subnetworks has yielded a large number of graph theoretical ways for identifying cohesive subgroups at the level of overall network structure. Limiting the discussion to one-mode networks, that is networks in which there can be a tie between any pair of vertices (for two-mode or bipartite networks, ▶ Social Network Analysis, Two-Mode Concepts in), Wasserman and Faust (see pp. 251–252 in [36]) distinguish 4 approaches to defining cohesive subgroups.

In the first and strictest approach, a cohesive subgroup is defined as a set of vertices in which all vertices are adjacent, that is, directly linked, to one another. In other words, cohesive subgroups are maximal complete subgraphs, which are called cliques [3].

The second approach is based on the notion of reachability and closeness of members within a subgroup. Members of a subgroup must be reachable in the sense that there are paths between them, i. e., a sequence of lines such that the end vertex of one line is the starting vertex of the next line (following the direction of the lines if they are arcs) such that no vertex in between the first and last vertex occurs more than once. In addition, the shorter the geodesics (shortest paths) between them, the closer the vertices are in a graph theoretical sense, so the more cohesive the subgroup is supposed to be.

The criterion of reachability does not necessarily yield very dense subgroups. In sparse networks, any maximal connected subgraph (strong component) may represent a cohesive subgroup: There is a path between each pair of vertices within a component. Increasing the required number of independent paths between any pair of vertices within a cohesive subgroup yields slightly denser subgroups, e. g., requiring at least two independent paths produces bi-components, which may be generalized to k-components for higher minimum numbers of independent (vertex-disjoint) paths between all vertices within a subgroup [37].

Focusing on graph-theoretical distance between vertices usually yields denser subgroups. One may, for instance, set a maximum n to the distance between any two vertices within a subgroup, which is the concept of an n-clique [38,39]. Adding the restriction that the diameter of an n-clique is n or less, one obtains n-clans [38,40]. Alternatively, one may define a cohesive subgroup as a maximal subgraph of diameter n, which is called an n-club [40].

The third approach focuses on the minimum number, strength, or multiplicity of ties among subgroup members. Subgraphs that are maximal with respect to the minimum number of neighbors within the subgraph are called k-cores [41] and a maximal subgraph with respect to the maximum number of vertices in the subgraph that are not adjacent are known as k-plex [42]. In a similar vein, restrictions can be imposed on the minimum strength or multiplicity of ties among members of a cohesive subgroup, generalizing Seidman's concept of a k-core to a valued core, which is called an m-core (see pp. 115–116 in [43]) or m-slice (see pp. 109–110 in[44]): Maximal connected subgraphs considering only lines with minimum value (or multiplicity) m.

In the fourth approach, cohesive subgroups are based on the relative frequency of ties among subgroup members in comparison to non-members: Cohesive subgroups are relatively dense sections within the network, that is, relative to the sections outside (and between) subgroups. An *LS* set [45] is a maximal subgraph such that any of its subsets has more ties to its complement within the *LS* set than to vertices outside the *LS* set. Borgatti, Everett and Shirey [46] generalized this idea to the concept of the lambda set, which requires the number of edge-disjoint paths between any pair of vertices within the lambda set to be larger than between any vertex within and any vertex outside the lambda set. A probabilistic version of plus-clusters in signed networks, discussed in the Introduction, can also be subsumed under this approach as it requires relatively many positive lines within cohesive subgroups and relatively many negative lines among cohesive subgroups [47]. Finally, clustering techniques and some types of blockmodels (see pp. 741–742 in [14]) also detect clusters of vertices that have relatively many ties within clusters and few among clusters. These models offer an alternative way for finding cohesive subgroups (see pp. 133–246 in [1]), ▶ Positional Analysis and Blockmodeling.

The large number of alternative measures for cohesive subgroups attests to the fact that behavioral tendencies at the actor level do not play out into nicely structured overall networks in a standard way. Especially the density of the social relation under investigation has an impact on the extent and ways in which cohesive subgroups can be found in the overall structure of the network. As a consequence, measures of network cohesion such as the clustering coefficient (averaged over all vertices in the network) can be quite uninformative about overall network structure: Loosely knit cohesive subgroups which are clearly identified by some of the techniques presented above, yield low clustering coefficients in a sparse network. Cohesion in a network is better summarized by calculating the percentage of vertices that are part of identified subgroups, the number and sizes of cohesive subgroups, and so on [48].

In addition to homophily, there is a second behavioral hypothesis related to cohesion in SNA. This hypothesis is based on the idea that social action is embedded in networks [49,50,51]. Named after the sociologist Georg Simmel, Simmelian ties are ties that are embedded in other ties, e. g., business ties are embedded in family ties, or in complete triads and cliques. They are hypothesized to enforce group norms and enhance trust, hence pressure people into the same behavior because the two actors involved in a tie share common neighbors who supervise their behavior.

Just as with the homophily hypothesis, the embeddedness hypothesis predicts that tightly connected actors will be more similar in their behavior and attitudes. In addition, it predicts that embedded ties are more stable and new ties are more likely to be established when they are embedded in existing cliques or existing ties. Closure again is an important indicator of tendencies to establish and maintain embedded ties but so is the multiplicity of relations: The extent to which a tie on one social relation duplicates a tie on another social relation. At the level of overall network structure, we should expect relatively dense sections, especially cliques, and in a multirelational network, that is, a network containing ties on two or more social relations, we should find that the same subsets of vertices are clustered on each relation. Graph-theoretical measures of the latter are rare. The stochastic blockmodeling technique developed by Krysztof Nowicki and Tom A.B. Snijders [52] is an example. See ► Social Networks, Algebraic Models for.

If data on vertex attributes are available, especially if they concern public behavior, that is, behavior that is easily noticed by third parties such as publicly expressed opinions and statements, Simmelian ties are hypothesized to produce a special effect. Involvement in different groups (cliques) then exposes actors to possibly conflicting sets of norms and loyalties, which may urge them to cut their ties with some or all of these groups [53]. In this case, actors are hypothesized to withdraw from stressful relations, so they discontinue ties that incorporate them into cliques (with a preference for cutting a minimum number of ties) or they discontinue ties such that they are no longer connected to actors voicing different opinions or norms. At the macro level, this would produce disconnected sets of cliques instead of overlapping cliques.

Centrality and Brokerage

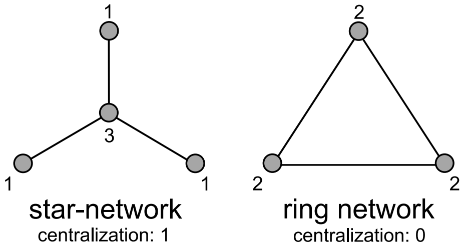
The notion of centrality in social networks has a long history in SNA. It is attributed to Alex Bavelas [54]. In discussions of centrality, network ties are usually regarded as channels for the exchange of information, goods, services, and so on. Being central in this exchange system has always been hypothesized to be related to influence and satisfaction. Centralization, as a characteristic of a network, has been linked to the efficiency of a network as an exchange system. More centralized groups, for example, have often been shown to be more efficient.

Linton C. Freeman [55] argued that the approaches to centrality are based on three ideas about what being central means: (1) being active within the network, that is, maintaining many ties, (2) being efficient or independent of go-betweens by having short distances to other vertices in the network, and (3) being an important go-between, that is, being part of many paths between other vertices in the network. Although alternative classifications and approaches exist, for instance, Noah E. Friedkin's alternative classification [56] and the formal graph theoretical approach to centrality by Stephen P. Borgatti and Martin G. Everett [57], Freeman's classification is used here, adding concepts of brokerage that have been developed elsewhere in SNA.

Activity

Being active or prominent in the network means that an actor has many ties, hence access to many sources of information (etc.). As a consequence, this actor is more attractive as a neighbor for other actors, which translates to the behavioral hypothesis that actors have a preference for ties with vertices that already have many ties. The degree of a vertex (the number of lines incident with a vertex), then, is the relevant graph theoretical measure of local structure, which is also known as degree centrality. Note that the concept of centrality expresses a structural property of a vertex.

Centralization is the corresponding structural property of a network and it is defined as the variation in the centrality scores of the vertices in the network because this variation shows the extent to which there is a center (very central vertices) and a periphery (vertices with very low centrality scores). The star and ring networks are defined



Social Network Analysis, Graph Theoretical Approaches to, Figure 4

Centrality and centralization in a star network and a ring network

as respectively the most and least centralized networks and they are known to exhibit the highest and lowest variation in centrality scores in simple networks, that is, networks without multiple lines (and loops in the case of a directed network). See Fig. 4 for an illustration, showing a star and a ring network labeling the vertices with their degree centrality scores.

From research on the power law in networks [58,59], discussed in another entry of this encyclopedia, it is known that preferential attachment to degree creates networks with a peculiar degree distribution, including many vertices with low or modest degree and few vertices with high degree. According to the definition of centralization in network analysis, this implies large variation in degree centrality scores, hence high degree centralization. In other words, the behavioral hypothesis of preference for high degree actors produces centralized overall network structure.

Note, however, that the way in which the 'power law networks' are assembled – growing from an initial seed without context – is hardly ever applicable to social networks, which usually have no discernable starting point (networks originate from networks) and are always constrained by the historical context. It remains to be seen whether power law distributions in empirical social networks are created by preferential attachment. Some results indicate that even though the degree distributions of cross-sectional snapshots of a large social network follow the power law, there is hardly any continuity in degree centrality of vertices over time, which does not suggest that the actors are driven by preferential attachment [60].

Efficiency and Weak Ties

The second approach to centrality focuses on graph theoretical distances between vertices. The central idea here is that actors try to improve access to information and efficient spreading of information by minimizing the number of go-betweens needed to reach or be reached by all other actors in the network. The behavioral hypothesis states that a vertex prefers to link to actors that give access to parts of the network that are presently remote to this vertex and that can only be reached through some or many go-betweens that may withhold or distort information. A minimum number of go-betweens yields maximum independence and maximum efficiency in the exchange network.

Graph theoretical measures of local structure focus on graph theoretical distance, that is, the minimum length of paths between vertices because path length equals the number of go-betweens in the network plus one. Linton Freeman's closeness centrality [55] is a straightforward implementation of this idea because it merely normalizes the average graph theoretical distance between a vertex and all other reachable vertices in the network. In addition, paths can be weighted by the centrality of the vertices on them, which is done by Phillip Bonacich's eigenvector centrality [61,62]. Finally, the difference between incoming paths and outgoing paths may be added, which is implemented in the version of closeness centrality developed by Thomas W. Valente and Robert K. Foreman [63].

Again, the normalized variation of closeness centrality scores of the vertices in the network yields the appropriate measure of centralization of overall network structure. It is not known yet, however, whether and how tendencies to reduce distances to other vertices at the level of individual actors in the network play out into the closeness centralization of the overall network. On the one hand, if the network contains some vertices with high closeness centrality, they offer rather short paths toward many vertices so they should attract a lot of new ties. This would enhance their centrality and possibly the variation in centrality scores although a general rise in closeness centrality scores of all vertices may also decrease the variation. On the other hand, if the network has low centralization, it is more likely that vertices connect directly to remote parts, reducing path lengths among remote parts, which would not raise the variation in closeness centrality scores and yield or sustain low closeness centralization of overall network structure.

The strength of weak ties argument proposed by Mark Granovetter [64,65] may be regarded as a special application of the notion of efficiency. In his research on finding a job, Granovetter noticed that relatively superficial ties, ties with infrequent contact, give access to new information because they are more likely to link you to someone with whom you are not linked directly or at a short distance. Strong or intense ties tend to be situated within cohesive subgroups, so they are more likely to offer redundant information already received through other ties. Granovetter is only interested in the effects of having weak ties, but if we turn his idea into a behavioral hypothesis, it suggests a preference to relate to distant vertices that are neither connected to yourself nor to your neighbors (or your neighbors' neighbors, and so on).

A tendency to connect to the most remote parts of the network means that actors tend to establish links to vertices at a large or maximum distance in the network. As in the case of maximizing closeness centrality, it is not clear whether this leads to identifiable patterns in overall network structure. It is quite obvious, however, that this tendency acts as a counterforce against tendencies toward cohesion as densely connected subnetworks: Actors are hypothesized to span gaps rather than to close local configurations. Different hypotheses must be developed for strong ties, which are hypothesized to contribute to subgroup formation, and weak ties linking remote parts. The weak ties will probably increase the number of links between dense parts of the network, increasing the *k*-connectivity (minimum number of node-independent paths between any two vertices) of the network. If so, we should expect high *k*-connectivity of the network if the weak ties hypothesis is true [37].

Control and Structural Holes

The third approach to centrality focuses on control over flows within the network: The more you are in between other vertices in the network, the more they depend on you to pass on information, the more you are able to control exchange within the network and profit from your control. Using this type of control is called brokerage.

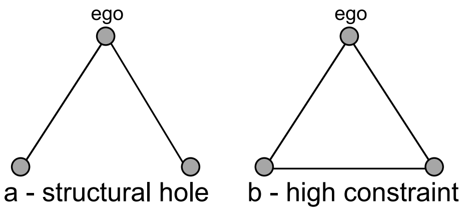
The notion of being in between other vertices has a straightforward translation to graph theory as being part of a path between two other vertices. Limiting paths to the shortest paths between vertices both Linton C. Freeman's [55] betweenness centrality and Jac M. Anthonisse's [66] rush of a vertex are based on the proportion of all geodesics between other vertices that include this vertex. This measure has been extended to handle directed ties [67,68]. Information centrality [69,70] takes into account all paths between vertices, not just the geodesics and flow betweenness [71] or entropy [72] also consider the values of lines. See Stephen P. Borgatti's [73] classification for more details.

Betweenness centralization as the normalized variation of betweenness centrality of the vertices in the network offers a measure of centralization at the network level and so do center-periphery blockmodels (see pp. 741–742 in [14]), e. g., in the world trade system, central countries are able to profit from the lack of trade among countries in the periphery [74,75,76,77]. The link between strategies for maximizing betweenness centrality by actors and the overall structure of the resulting network is unclear. It is not to be expected that control behavior will produce a highly centralized structure because no actor will be satisfied with low betweenness centrality and as a consequence high betweenness centrality for some vertices is unlikely. Therefore, the variation in betweenness centrality scores will not be high.

While betweenness centrality situates an actor with respect to all other actors in the network, Ronald S. Burt [78,79] proposed a local variant, focusing on control within the ego-network of an actor. His behavioral hypothesis rests on the *tertius gaudens* principle: The benefits that accrue to an actor that is in between two actors that are not directly linked because of the opportunity to broker information between them or, in a more malicious variant, to divide and conquer.

This hypothesis translates quite easily into graph theoretical structure. The absence of a tie between two neighbors of an actor is called a structural hole (Fig. 5a). The behavioral hypothesis states that actors try to increase the structural holes that they can exploit. At the same time, however, they try to minimize the structural holes through which they can be exploited. This means, among other things, that an actor will not end a tie to one of its neighbors if the two neighbors are directly linked (see Fig. 5b): That would create an opportunity to broker at the expense of the actor. In this situation, the actor is constrained in its opportunities to change ties.

Structural holes and constraint are the flip sides of the same coin. A tie with low constraint indicates that the tie is involved in (many) incomplete triads (such as Fig. 5a), so there are (many) structural holes offering the actor options for brokering. High constraint on a tie means that it is part of (many) complete triads (as in Fig. 5b), so there are few or no possibilities for brokerage. Because the presence or absence of ties among an actor's neighbors is key



Social Network Analysis, Graph Theoretical Approaches to, Figure 5
A structural hole (a) and a triad with high constraint (b)

to the argument, network analysts have also used the density of the ego-network without the ego as a proxy of constraint: The higher the density, the higher the constraint on the ego. Alternatively, betweenness centrality for ego-networks [80,81] can be used.

If the structural holes hypothesis governs actors' behavior, what overall network structure should we expect to find? A strong tendency toward brokerage at the micro level is not likely to produce a centralized network because each actor would try to maximize the number of structural holes around itself, which would yield a bipartite graph in the extreme case (no ties among any vertex' neighbors), or it would minimize its constraint by ending all ties to neighbors that have contacts outside ego's immediate neighborhood, which would produce a highly clustered network consisting of isolated cliques or isolated vertices in the extreme case.

In summary, the relation between local action and overall network structure is simple and clear only in the case of preferential attachment to high-degree vertices. Tendencies to maximize centrality that looks beyond the immediate neighbors such as closeness and betweenness centrality, do not necessarily yield centralized networks. Even for local structures, alternative hypotheses for actor's behavior are available that are unlikely to produce centralized networks, e. g., a preference to avoid constraint. The interplay between local action and overall structure is quite complex.

Prestige and Ranking

The preceding sections have not distinguished between directed and undirected networks. For prestige and ranking, however, the direction of ties is crucial because asymmetry in networks is assumed to be linked to social prestige [82]. The general idea here is that social inequalities are reflected and possibly created by asymmetric ties, e. g., everyone invites the most popular boy or girl in class but s/he doesn't return each invitation. Of course, the nature of the social relation determines the direction of choices; ties like "reports to" or "pays respect to" point toward higher levels in a hierarchy while "beating up" points in the opposite direction.

A central behavioral hypothesis concerns the popularity or attractiveness of actors. Actors tend to (want to) relate to actors with attractive structural properties or attributes, so attributes related to power or social status increase the probability that an actor will be chosen. From a constructivist point of view, however, being chosen often is also interpreted as a sign of importance and prestige, so receiving many choices (ties) increases the probability of receiving even more. In this way, networks may produce informal status hierarchies. The Matthew Effect, proposed by Robert K. Merton [83], comes to mind here: "For unto every one that hath shall be given, and he shall have abundance: But for him that hath not shall be taken away even that which he hath" (gospel of Matthew XXV, 29).

In graph theoretical terms, the structural attractiveness of an actor refers to the number of incoming arcs on vertices, which is simply the indegree of a vertex. This is called the popularity of a vertex and, of course, we must replace it by the vertex' outdegree if the relation is negative, e. g., submission, beating up, criticizing. If indirect choices must be taken into account as well, attractiveness is measured by proximity prestige [84], which is based on the average distance from all other vertices in the network – a directed variant of closeness centrality. Proximity prestige captures the idea that nominations or choices by actors who are themselves popular, contribute more to ones structural prestige. Bonacich's measure of power [62,85] adds the idea that power may also be derived from being connected to powerless actors rather than to other powerful people.

The popularity hypothesis is another example of preferential attachment to degree with the restriction that we focus either on indegree or outdegree. Therefore, the indegree (or outdegree) distribution of the overall network is expected to follow the power law and network structure will be characterized by high degree centralization.

Adding data on social attributes that make some actors more prestigious such as wealth, social class, beauty, and so on, we should expect preferential attachment to vertices that score high on these attributes. Note that vertex attributes play a slightly different role here than in the case of cohesion. Now we are concerned with attributes of (at least) ordinal level, expressing prestige that an actor possesses to a higher or lower degree. In the case of cohesion, we deal with nominal attributes, which merely express an identity. In contrast to homophily, the attributes of the actor who initiates the directed tie (the tail of the arc) does not matter here because the effect is solely related to structural characteristics or attributes of the actor at the receiving end of the tie (the arc's head).

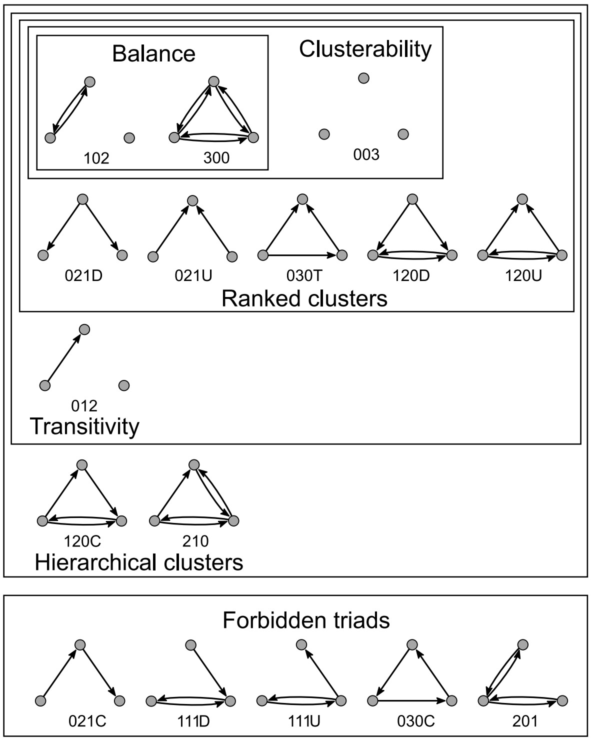
There is a second, slightly different behavioral hypothesis relating to deference or submission rather than attractiveness. The idea is that actors mainly tend to create positive ties to other actors in their own status group or to actors in a higher status group – the people they are looking up to – to consolidate and improve their social position. Similarly, they tend to direct negative ties to actors in lower status groups. Note the difference with attractiveness: It is hypothesized that actors choose upwardly but they need not prefer the most attractive (top) actors in the network as they are supposed to do according to the attractiveness hypothesis.

The main difference between attractiveness and deference is that the former only takes into account structural properties or attributes of the tie's receiver, whereas the characteristics of both the sender and receiver of the tie matter to the latter. This distinction has important consequences to the structure of the overall network. Whereas the attractiveness hypothesis yields centralized networks, the deference hypothesis yields layered networks. The layers consist of sets of vertices that are symmetrically linked, e. g., by reciprocal ties, while the ties between layers are asymmetric, all pointing in the same direction. This behavior may be both a consequence of a formal hierarchy, e. g., positions within an organization with formalized relations such as reports to, or actually show an informal social hierarchy, e. g., status differences between men and women in a particular social setting.

In simple directed networks, triads, that is, three vertices and the lines among them, are the key to measuring tendencies toward ranking at the local level. Translating the concepts of balance and clusterability (see Sect. "Introduction") from signed digraphs to unsigned digraphs, James A. Davis and Samuel Leinhardt [86,87] replaced positive ties by symmetric ties within cohesive subgroups and negative ties by no ties among subgroups. Thus, ties within a layer, either within clusters or among clusters, are symmetric. They assumed that asymmetric ties represent the ranking of clusters into a hierarchy, introducing the model of ranked clusters. Moreover, they showed that a network with a perfectly fitting ranked clusters model contains only certain types of triads, whereas other types do not occur (Fig. 6).

The ranked clusters model requires arcs from each vertex to all vertices on higher ranks. This requirement is usually too strict for empirical social networks and it is relaxed in the transitivity model proposed by Paul W. Holland and Samuel Leinhardt [88], which requires that clusters of vertices on different ranks are either completely linked or not linked at all, yielding a partial order, by simply adding one type of triad to the set of allowed triads (Fig. 6). Later, Eugene Johnsen [89] proposed the model of hierarchical clusters to account for asymmetries within clusters. Note the nesting of the models for overall network structure, which is why the sets of permitted triads is extended for more general models. Finally, it was shown that the models developed for unsigned digraphs could also be detected in incomplete signed digraphs using types of semicycles [90].

The triads characterizing ranked structures serve as models for tie creation, maintaining, and breaking behavior of actors under the deference hypothesis. For instance, triad 120U (Fig. 6) predicts that a member of a cluster is likely to establish or maintain a tie to an actor at a higher rank if its neighbors within the cluster have such a tie. In the perfect case, there is a one-to-one relation between sets of occurring types of triads and the overall structure of the network. Therefore, triad census [91,92], which is the frequency distribution of the sixteen types of triads in a directed network, offers an indication of overall network structure. In the imperfect case, the triad census of a network may be compared to the frequency distributions of triad types in randomly generated networks to test the tendency toward ranking.



Social Network Analysis, Graph Theoretical Approaches to, Figure 6
Triad types and balance-theoretic models

The triad census does not show the composition of the clusters and ranks; it does not identify the vertices belonging to particular clusters and ranks. This can be done in several ways. Realizing that ranks should be connected asymmetrically in directed networks, strong components cannot include more than one rank because vertices within strong components are by definition mutually reachable. Ties between strong components, then, are asymmetric, so it is easy to establish the ranking among strong components. Strong components, however, do not require a lot of symmetry in the ties; actually, no tie needs to be reciprocated. The symmetric-acyclic decomposition proposed by Patrick Doreian, Vladimir Batagelj and Anuška Ferligoj [93] does require at least some symmetric ties (mutual choices) within clusters because they define a symmetric cluster as a maximum subset of vertices that are directly or indirectly linked by symmetric ties. Generalized blockmodels [1] that are asymmetric with respect to off-diagonal blocks offer another way to identify hierarchical relations (see ► Positional Analysis and Blockmodeling).

Analyzing Complexity in Social Networks

The preceding sections presented behavioral hypotheses that have similar, different, or even opposite consequences for overall network structure. It is not plausible that one particular type of behavior dominates network formation. Therefore, it is not likely that overall structure of empirical social networks will display one particular form that can be hypothesized in advance or that behavioral tendencies can be adequately tested on particular characteristics of overall network structure. It has been shown, for example, that the degree distribution of a network does not reveal tendencies toward cohesive subgroup formation that are operative during network evolution [94].

Even if overall network structure displays certain characteristics, they may be produced by different types of behavior. Centralization in a social network, for instance, may arise from a tendency of actors to minimize paths to all other actors or from a preference for prestigious actors. Alternatively, it may be a by-product of cohesive subgroup formation: Actors that are marginal to the cohesive subgroups may directly or indirectly connect different subgroups, which gives them a central position with respect to betweenness. Furthermore, pronounced overall network patterns may occur only temporarily in empirical social networks when they create socially unstable situations. The polarization predicted by balance theory in Sampson's network of novices, used as an example in the Introduction, was only temporary. After this polarization and most likely due to it, many novices left the monastery. The network, so to speak, fell apart.

For these reasons, SNA increasingly focuses on local structure using overall network structure merely as a collection of (overlapping) local structures. Behavioral hypotheses translate much more directly to local structure, that is, to the ties of the actor and those of its neighbors (and possibly their neighbors), as we have seen in the preceding sections. Local structure is the part of the network that an actor can easily survey and actually change.

The latest developments in techniques for modeling network structure and evolution apply this actor-oriented approach either in statistical models, see ► Network Analysis, Longitudinal Methods of and ► Social Networks, Exponential Random Graph (p^*) Models for. The techniques test behavioral tendencies by relating the creation, maintenance, and ending of ties by individual actors to the local configuration of ties, to previous ties between the actor and the alter or to present ties on another social relation, and to characteristics of both the actor and the alter.

In principle, the actor-oriented approach is able to test all behavioral hypotheses presented in the preceding sections on characteristics of local structure in which the actors are embedded and properties of the actors themselves. If hypothesized local configurations appear more often than expected by chance, the underlying behavioral hypothesis is assumed to guide individual behavior at least to some extent. If the behavior in a set of actors or individual actor's behavior are in line with several behavioral hypotheses at the same time, the effect of each behavioral tendency can be separated. Thus, it is possible to link complex overall network structure to compound behavior of the actors in the network.

Future Directions

The techniques for analyzing local structure are in development. Models for the co-evolution of relations and quantitative attributes of vertices over time have just been introduced [95]. Not all behavioral hypotheses have been included yet, for instance, because they involve non-standard types of networks such as signed relations, and new ones are bound to be proposed. Incomplete data and external constraints on data collection or conceptual constraints on network structure such as two-mode networks (see ► Social Network Analysis, Two-Mode Concepts in) may limit the applicability of current models and spur the development of new ones.

If the actor-oriented approach is successful, will overall network structure become completely redundant? Will it only serve network exploration – looking for behavioral hypotheses rather than testing them – and for analyzing the consequences of network position on behavior, attitudes, or esteem, for instance, does the subgroup to which an actor belongs or its centrality correlate with its subsequent behavior or attitudes?

Let us return to the balance example, presented in the Introduction. The high degree of polarization among Sampson's novices was followed by the voluntary or forced exit of several novices. This suggests that people are able to survey overall network structure and draw conclusions from it, rather than just react to their local network environment. A highly polarized social group does not seem to be the kind of situation that we like to be living in. An actor's perception of overall network structure may also affect its behavior.

In this respect, the balance example is interesting because the original theory by Heider referred to perceptions of affect relations and unit relations (having characteristics or possessing items) rather than objectively measured relations. The importance of perception to network analysis is stressed in a more general way by Harrison White, who argues that social ties are stories [96]. According to him, people are linked into social networks by the stories that they tell about their ties. Thus, we should expect people to react to the ties as they remember and tell them rather than to the ties as they are observed by the researcher or registered in, for instance, membership lists.

A similar argument was made by David Krackhardt when he reconsidered balance theory [97]. According to him, we should measure each actor's perception of network structure, for which he proposed the concept of Cognitive Social Structures, compare these perceptions, and use them to explain why actors behave as they do. In his approach, actors are assumed to be able and active in forming impressions of overall network structure.

If humans are capable of imagining overall network structure, then communication of these perceptions may also play a role in network formation, including the accidental and deliberate distortions or simplifications that are likely to happen. This brings us to the link between network structure and mental categories such as social classifications or culture in a more general sense as argued by Ronald L. Breiger [98,99]. If, for example, members of the network perceive and discuss cohesive subgroups, they assign names and meanings to social configurations. Thus, social meanings and identities are created in the process of establishing social ties and interpreting them. The duality of social structure on the one hand and the structure of symbolic categories on the other hand as proposed by John W. Mohr [100], may be the essential condition for cultural meanings that are social in the sense that they affect actors' sense of identity and behavior.

These discrete and qualitative rather than continuous and quantitative classifications are very likely to affect the network behavior of actors: They define the categories that are experienced as being similar or dissimilar in the case of homophily and group formation, or superior versus inferior in the case of prestige and ranking. As a consequence, it is to be expected that the present focus on local structure will be complemented by a focus on overall network structure, especially perceived and communicated network structure.

Bibliography

Primary Literature

1. Doreian P, Batagelj V, Ferligoj A (2005) Generalized blockmodeling. Cambridge University Press, Cambridge
2. Freeman LC (2004) The development of social network analysis. A study in the sociology of science. Empirical, Vancouver
3. Luce RD, Perry A (1949) A method of matrix analysis of group structure. Psychometrika 14:95–116
4. Cartwright D, Harary F (1956) Structural balance: A generalisation of Heider's theory. Psychol Rev 63:277–293
5. Murphy G (1937) Editorial foreword. Sociometry 1:5–7
6. Moreno JL (1937) Sociometry in relation to other social sciences. Sociometry 1:206–219
7. Warner WL, Lunt PS (1941) The social life of a modern community. Yale University Press, New Haven
8. Davis A, Gardner BB, Gardner MR (1946) Deep south: A social anthropological study of caste and class. University of Chicago, Chicago
9. Mitchell JC (1969) The concept and use of social networks. In: Mitchell JC (ed) Social networks in urban settings. Manchester University Press, Manchester
10. Ore Ø (1963) Graphs and their uses. Mathematical Association of America, Washington
11. Flament C (1963) Applications of graph theory to group structure. Prentice-Hall, Englewood Cliffs
12. Harary F, Norman RZ, Cartwright D (1965) Structural models: An introduction to the theory of directed graphs. Wiley, New York
13. Rogers EM (1962) Diffusion of innovations. Free Press, New York
14. White HC, Boorman SA, Breiger RL (1976) Social structure from multiple networks. I. Blockmodels of roles and positions. Am J Sociol 81:730–780
15. Helmers HM, Mokken RJ, Plijter RC, Stokman FN, Anthonisse JM (1975) Graven naar macht: Op zoek naar de kern van de Nederlandse economie. Van Gennep, Amsterdam
16. Wellman B (1988) Structural analysis: From method and metaphor to theory and substance. In: Wellman B, Berkowitz SD (eds) Social structures: A network approach. Cambridge University Press, Cambridge, pp 19–61
17. Simmel G (1902) The number of members as determining the sociological form of the group. I. Am J Sociol 8:1–46
18. Heider F (1946) Attitudes and cognitive organization. J Psychol 21:107–112
19. Davis JA (1967) Clustering and structural balance in graphs. Hum Relat 20:181–187
20. Sampson SF (1968) A novitiate in a period of change. An experimental and case study of social relationships. Cornell University, Ithaca, p 575
21. Homans GC (1950) The human group. Harcourt Brace, New York
22. Moreno JL (1953) Who shall survive?: Foundations of sociometry, group psychotherapy and sociodrama, 1st edn 1934. Beacon House, Beacon
23. Lazarsfeld PF, Merton RK (1954) Friendship as a social process: A substantive and methodological analysis. In: Berger M, Abel

T, Page CH (eds) Freedom and control in modern society. Van Nostrand, Princeton, pp 18–66

24. McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: Homophily in social networks. Ann Rev Sociol 27:415–444
25. Festinger L (1962) A theory of cognitive dissonance, 1st edn 1957. Tavistock, London
26. Heise DR (1979) Understanding events: Affect and the construction of social action. Cambridge University Press, New York
27. Smith-Lovin L, Heise DR (1988) Analyzing social interaction: Advances in affect control theory. Gordon Breach Science, New York, pp 192
28. Morris M (1991) A log-linear modeling framework for selective mixing. Math Biosci 107:349–377
29. Newman MEJ (2003) Mixing patterns in networks. Phys Rev E 67
30. Tallis GM (1985) Transfer systems and covariance under assortative mating. Theor Appl Genet 70:497–504
31. Robins G, Elliott P, Pattison P (2001) Network models for social selection processes. Soc Netw 23:1–30
32. Robins G, Pattison P, Elliott P (2001) Network models for social influence processes. Psychometrika 66:161–189
33. Friedkin NE, Johnsen EC (1990) Social influence and opinions. J Math Sociol 15:193–205
34. Watts DJ (1999) Small worlds. The dynamics of networks between order and randomness. Princeton University Press, Princeton
35. Blau PM (1977) Inequality and heterogeneity: A primitive theory of social structure. Free Press, New York
36. Wasserman S, Faust K (1994) Social network analysis: Methods and applications. Cambridge University Press, Cambridge
37. White DR, Harary F (2001) The cohesiveness of blocks in social networks: Node connectivity and conditional density. Sociol Methodol 31:305–359
38. Alba RD (1973) A graph-theoretical definition of a sociometric clique. J Math Sociol 3:113–126
39. Luce RD (1950) Connectivity and generalized cliques in sociometric group structure. Psychometrika 15:169–190
40. Mokken RJ (1979) Cliques, clubs and clans. Qual Quant 13:161–173
41. Seidman SB (1983) Network structure and minimum degree. Soc Netw 5:269–287
42. Seidman SB, Foster BL (1978) A graph-theoretic generalization of the clique concept. J Math Sociol 6:139–154
43. Scott J (1991) Social network analysis: A handbook. Sage, London
44. de Nooy W, Mrvar A, Batagelj V (2005) Exploratory social network analysis with Pajek. Cambridge University Press, Cambridge
45. Seidman SB (1983) Internal cohesion of ls sets in graphs. Soc Netw 5:97–107
46. Borgatti SP, Everett MG, Shirey PR (1990) LS sets, lambda sets and other cohesive subsets. Soc Netw 12:337–357
47. Doreian P, Mrvar A (1996) A partitioning approach to structural balance. Soc Netw 18:149–168
48. White DR, Owen-Smith J, Moody J, Powell WW (2004) Networks, fields and organizations: Micro-dynamics, scale and cohesive embeddings. Comput Math Organ Theory 10:95–117
49. Friedkin NE (1984) Structural cohesion and equivalance explanations of social homogeneity. Sociol Methods Res 12:235–261
50. Granovetter M (1985) Economic action and social structure: The problem of embeddedness. Am J Sociol 91:481–510
51. Granovetter M (1992) Problems of explanation in economic sociology. In: Nohria N, Eccles RG (eds) Networks and organizations: Structure, form, and action. Harvard Business School Press, Boston, pp 25–56
52. Nowicki K, Snijders TAB (2001) Estimation and prediction for stochastic blockstructures. J Am Statist Assoc 96:1077–1087
53. Krackhardt D (1999) The ties that torture: Simmelian tie analysis in organizations. Res Sociol Organ 16:183–210
54. Bavelas A (1948) A mathematical model for group structures. Hum Organ 7:16–30
55. Freeman LC (1979) Centrality in social networks conceptual clarification. Soc Netw 1:215–239
56. Friedkin NE (1991) Theoretical foundations for centrality measures. Am J Sociol 96:1478–1504
57. Borgatti SP, Everett MG (2006) A graph-theoretic perspective on centrality. Soc Netw 28:466–484
58. Albert R, Barabasi A-L (2000) Topology of evolving networks: Local events and universality. Phys Rev Lett 85:5234–5237
59. Barabási A-L (2002) Linked: The new science of network. Perseus, Cambridge
60. Braha D, Bar-Yam Y (2006) From centrality to temporary fame: Dynamic centrality in complex networks. Complexity 12:59–63
61. Bonacich P (1972) Factoring and weighting approaches to clique identification. J Math Sociol 2:113–120
62. Bonacich P (1987) Power and centrality: A family of measures. Am J Sociol 92:1170–1182
63. Valente TW, Foreman RK (1998) Integration and radiality: Measuring the extent of an individual's connectedness and reachability in a network. Soc Netw 20:89–105
64. Granovetter M (1973) The strength of weak ties. Am J Sociol 78:1360–1380
65. Granovetter M (1995) Getting a job: A study of contacts and careers, 1st edn 1974. The University of Chicago Press, Chicago
66. Anthonisse JM (1971) The rush in a directed graph. Stichting Mathematisch Centrum, Amsterdam, p 10
67. Gould RV (1987) Measures of betweenness in non-symmetric networks. Soc Netw 9:277–282
68. White DR, Borgatti SP (1994) Betweenness centrality measures for directed graphs. Soc Netw 16:335–346
69. Stephenson K, Zelen M (1989) Rethinking centrality: Methods and examples. Soc Netw 11:1–37
70. Newman MEJ (2005) A measure of betweenness centrality based on random walks. Soc Netw 27:39–54
71. Freeman LC, Borgatti SP, White DR (1991) Centrality in valued graphs: A measure of betweenness based on network flow. Soc Netw 13:141–154
72. Tutzauer F (2007) Entropy as a measure of centrality in networks characterized by path-transfer flow. Soc Netw 29:249–265
73. Borgatti SP (2005) Centrality and network flow. Soc Netw 27:55–71
74. Snyder D, Kick E (1979) Structural position in the world system and economic growth 1955–70. A multiple network analysis of transnational interactions. Am J Sociol 84:1096–1126

75. Breiger RL (1981) Structures of economic interdependence among nations. In: Blau PM, Merton RK (eds) Continuities in structural inquiry. Sage, Newbury Park, pp 353–380
76. Bornschier V, Trezzini B (1997) Social stratification and mobility in the world system – Different approaches and recent research. Int Sociol 12:429–455
77. Borgatti SP, Everett MG (2000) Models of core/periphery structures. Soc Netw 21:375–395
78. Burt RS (1992) Structural holes: The social structure of competition. Harvard University Press, Cambridge/London
79. Burt RS (1992) The social structure of competition. In: Nohria N, Eccles RG (eds) Networks and organizations: Structure, form, and action. Harvard Business School Press, Boston, pp 57–91
80. Everett M, Borgatti SP (2005) Ego network betweenness. Soc Netw 27:31–38
81. Marsden PV (2002) Egocentric and sociocentric measures of network centrality. Soc Netw 24:407–422
82. Knoke D, Burt RS (1983) Prominence. In: Burt RS, Minor MJ (eds) Applied network analysis. A methodological introduction. Sage, Beverly Hills, pp 195–222
83. Merton RK (1968) The Matthew effect in science. Science 159:56–63
84. Lin N (1976) Foundations of social research. McGraw-Hill, New York
85. Bonacich P, Lloyd P (2001) Eigenvector-like measures of centrality for asymmetric relations. Soc Netw 23:191–201
86. Davis JA, Leinhardt S (1968) The structure of positive interpersonal relations in small groups. Annual Meeting of the American Sociological Association, Boston, August 1968
87. Davis JA, Leinhardt S (1972) The structure of positive interpersonal relations in small groups. In: Berger J (ed) Sociological theories in progress, vol 2. Houghton Mifflin, Boston, pp 218–251
88. Holland PW, Leinhardt S (1971) Transitivity in structural models of small groups. Comp Group Stud 2:107–124
89. Johnsen EC (1985) Network macrostructure models for the Davis–Leinhardt set of empirical sociomatrices. Soc Netw 7:203–224
90. de Nooy W (1999) The sign of affection: Balance-theoretic models and incomplete signed digraphs. Soc Netw 21:269–286
91. Davis JA (1979) The Davis/Holland/Leinhardt Studies: An overview. In: Holland PW, Leinhardt S (eds) Perspectives on social network research. Academic, New York, pp 51–62
92. Holland PW, Leinhardt S (1978) An omnibus test for social structure using triads. Sociol Methods Res 7:227–256
93. Doreian P, Batagelj V, Ferligoj A (2000) Symmetric-acyclic decompositions of networks. J Classif 17:3–28
94. Snijders TAB (2003) Accounting for degree distributions in empirical analysis of network dynamics. In: Breiger R, Carley K, Pattison P (eds) Dynamic social network modeling and analysis: Workshop summary and papers. National Academic, Washington DC, pp 146–161
95. Snijders TAB, Steglich CEG, Schweinberger M (2007) Modeling the co-evolution of networks and behavior. In: Montfort KV, Oud H, Satorra A (eds) Longitudinal models in the behavioral and related sciences. Erlbaum, Mahwah, pp 41–71
96. White HC (1992) Identity and control; A structural theory of social action. Princeton University Press, Princeton
97. Krackhardt D (1987) Cognitive social structures. Soc Netw 9:109–134
98. Breiger RL (2004) The analysis of social networks. In: Hardy MA, Bryman A (eds) Handbook of data analysis. Sage, London, pp 505–526
99. Bian Yj, Breiger RL, Davis D, Galaskiewicz J (2005) Occupation, class, and social networks in urban China. Social Forces 83:1443–1468
100. Mohr JW (1998) Measuring meaning structures. Ann Rev Sociol 24:345–370

Books and Reviews

Berkowitz SD (1982) An introduction to structural analysis: The network approach to social research. Butterworths, Toronto

Brandes U, Erlebach T (2005) Network analysis: Methodological foundations. Springer, Berlin

Breiger RL (2004) The analysis of social networks. In: Hardy MA, Bryman A (eds) Handbook of data analysis. Sage, London, pp 505–526

Carrington PJ, Scott J, Wasserman S (2005) Models and methods in social network analysis. Cambridge University Press, Cambridge

Degenne A, Forsé M (1999) Introducing social networks. Sage, London

de Nooy W, Mrvar A, Batagelj V (2005) Exploratory social network analysis with Pajek. Cambridge University Press, Cambridge

Freeman LC, White DR, Romney AK (1992) Research methods in social network analysis, 1st edn 1989. Transaction, New Brunswick, pp 530

Knoke D, Kuklinski JH (1982) Network analysis. Sage, Beverly Hills

Marsden PV, Lin N (1982) Social structure and network analysis. Sage, Beverly Hills

Scott J (1991) Social network analysis: A handbook. Sage, London

Wasserman S, Faust K (1994) Social network analysis: Methods and applications. Cambridge University Press, Cambridge

Wellman B, Berkowitz SD (1988) Social structures: A network approach. In: Granovetter M (ed) Structural analysis in the social sciences. Cambridge University Press, Cambridge, p 497

Social Network Analysis, Large-Scale

VLADIMIR BATAGELJ
University of Ljubljana, Ljubljana, Slovenia

Article Outline

Glossary

For the basic notions on graphs and networks see the Articlea Wouter de Nooy: ► Social Network Analysis, Graph Theoretical Approaches to.

Network consists of vertices linked by lines and additional data about vertices and/or lines.

Network decomposition identification of parts of network and their interconnections. Usually it is described by a partition of set of vertices or set of lines.

Time complexity of algorithm describes how the time needed to run the algorithm depends on the size of the input data.

Reduction of network a network obtained by shrinking each cluster from a given partition into a vertex.

Condensation a reduction for strong connectivity partition.

Cut a subnetwork of vertices/lines with values of selected property above given threshold.

Island a connected subnetwork of selected size of (locally) important, with respect to selected property, vertices/lines.

Pattern searching identification of all appearances of selected small subnetwork (pattern or fragment) in a given network.

Topological sort procedure to determine a compatible ordering in acyclic network.

Definition of the Subject

A *network* is based on two sets: a set of *vertices* (nodes), that represent the selected *units*, and a set of *lines* (links), that represent *ties* between units. Each line has two vertices as its *end-points*; if they are equal it is called a *loop*. Vertices and lines form a *graph*. A line can be *directed* – an *arc*, or *undirected* – an *edge*.

Additional data about vertices or lines are usually known – their *properties* (attributes). For example: name/label, type, value, position, … In general

$$\text{Network} = \text{Graph} + \text{Data}\,.$$

The data can be measured or computed.

Formally, a *network* $\mathcal{N} = (\mathcal{V}, \mathcal{L}, \mathcal{P}, \mathcal{W})$ consists of the following:

- A *graph* $\mathcal{G} = (\mathcal{V}, \mathcal{L})$, where $\mathcal{V}$ is the set of vertices and $\mathcal{L} = \mathcal{E} \cup \mathcal{A}$. $\mathcal{E} \cap \mathcal{A} = \emptyset$ is the set of lines. $\mathcal{A}$ is the set of *arcs* and $\mathcal{E}$ is the set of *edges*.
- $\mathcal{P}$ – set of *vertex value functions* or properties: $p\colon \mathcal{V} \to A$
- $\mathcal{W}$ – set of *line value functions* or weights: $w\colon \mathcal{L} \to B$

The size of a network/graph is expressed by two numbers: number of vertices $n = |\mathcal{V}|$ and number of lines $m = |\mathcal{L}|$. In a *simple undirected* graph (no parallel edges, no loops) $m \leq \frac{1}{2} n(n-1)$; and in a *simple directed* graph (no parallel arcs) $m \leq n^2$.

For a family of graphs $\mathbb{G}$, we define a *density* of graph $\mathcal{G}$ as $\gamma(\mathcal{G}) = \frac{m(\mathcal{G})}{m_{\max}(\mathbb{G})}$.

Introduction

Small networks (some tens of vertices) can be represented by a picture and analyzed by many algorithms (***UCINET***, ***NetMiner***). Also *middle size* networks (some hundreds of vertices), if they are not dense, can still be represented by a picture, but some analytical procedures can't be used.

Till 1990 most networks were small – they were collected by researchers using surveys, observations, archival records, etc. The advances in IT allowed one to create networks from the data already available in the computer(s) or by browsing on the Internet. *Large* networks became reality. Large networks are too big to be displayed in detail; special algorithms are needed for their analysis (Pajek). The availability of large data sets also provided incentives to the boost of theoretical research in (large) network analysis (not only in the social sciences).

A recent overview of social network analysis software is given in Huisman and Van Duijn [28].

Large Networks and Complexity of Algorithms

Large networks have several thousands or millions of vertices. The upper limit to their size is technologically dependent – they can be stored in computer's memory; otherwise we deal with a *huge* network (see Abello et al. [1]).

Large networks are usually sparse $m \ll n^2$; typically $m = O(n)$ or $m = O(n \log n)$, see Table 1.

A collection of large networks is available from Pajek's datasets.

The *time complexity* of an algorithm describes how the time needed to run the algorithm depends on the size of the input data. In computer science the problems for which only algorithms of exponential (or higher) complexity are known are considered hard or intractable, since the speed-up of a computer only additively increases the size of problems that can be solved in a given period of

Social Network Analysis, Large-Scale, Table 1
Examples of large networks

Network	$n = \lvert\mathcal{V}\rvert$	$m = \lvert\mathcal{L}\rvert$	Source
ODLIS dictionary	2909	18419	ODLIS online
Citations SOM	4470	12731	Garfield's collection
Molecula 1ATN	5020	5128	Brookhaven PDB
Comput geometry	7343	11898	BibTeX bibliographies
English words 2-8	52652	89038	Knuth's English words
Internet traceroutes	124651	207214	Internet Mapping Project
Franklin genealogy	203909	195650	Roperld.com gedcoms
World-Wide-Web	325729	1497135	Notre Dame Networks
Internet Movie DB	1324748	3792390	IMDB
Wikipedia	659388	16582425	Wikimedia
US patents	3774768	16522438	Nber
SI internet	5547916	62259968	Najdi Si

time, but the problems for which an algorithm of polynomial complexity exists are considered 'nice'. When dealing with large instances of problems this isn't always true anymore. Let us look to time complexities of some typical algorithms in Table 2.

For the interactive use on large networks already quadratic algorithms, $O(n^2)$, are too slow – we have to restrict our 'toolbox' to a selection of efficient, *subquadratic* algorithms.

How can we deal with large structures? Already Romans knew – ***divide et impera*** (divide and conquer). In case of networks ***divide*** means the use of (recursive) *decomposition* of a large network into several smaller networks (see Fig. 1) that can be visualized and treated further using more sophisticated methods; ***impera*** means that we have to take care about the interlinks among so obtained parts.

Another approach is the use of different *statistical* quantities to describe the properties of a network and using probabilistic models to derive the answers to some questions.

Decompositions

Decompositions of a network are usually described by clusterings of vertices or lines. In the following we shall use mainly the clusterings of vertices.

A nonempty subset $C \subseteq \mathcal{V}$ is called a *cluster* (group). A nonempty set of clusters $\mathbf{C} = \{C_i\}$ forms a *clustering*.

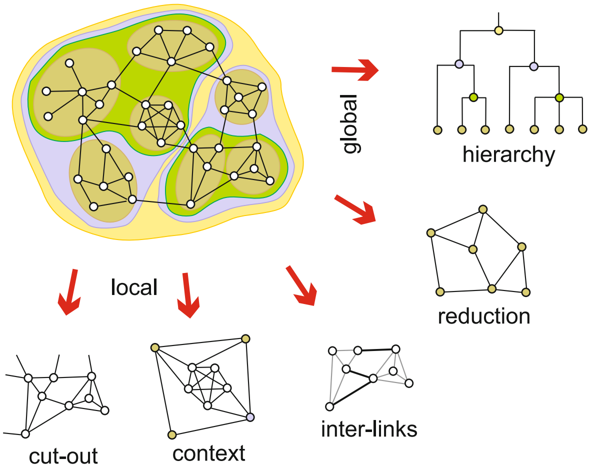
Clustering $\mathbf{C} = \{C_i\}$ is a *partition* iff

$$\cup\mathbf{C} = \bigcup_i C_i = \mathcal{V} \quad \text{and} \quad i \neq j \Rightarrow C_i \cap C_j = \emptyset\,.$$

Clustering $\mathbf{C} = \{C_i\}$ is a *hierarchy* iff $C_i \cap C_j \in \{\emptyset, C_i, C_j\}$. In other words, in a hierarchy two clusters are either disjoint or is one contained in the other.

Hierarchy $\mathbf{C} = \{C_i\}$ is *complete*, iff $\cup\mathbf{C} = \mathcal{V}$; and is *basic* if for all $v \in \cup\mathbf{C}$ also $\{v\} \in \mathbf{C}$.

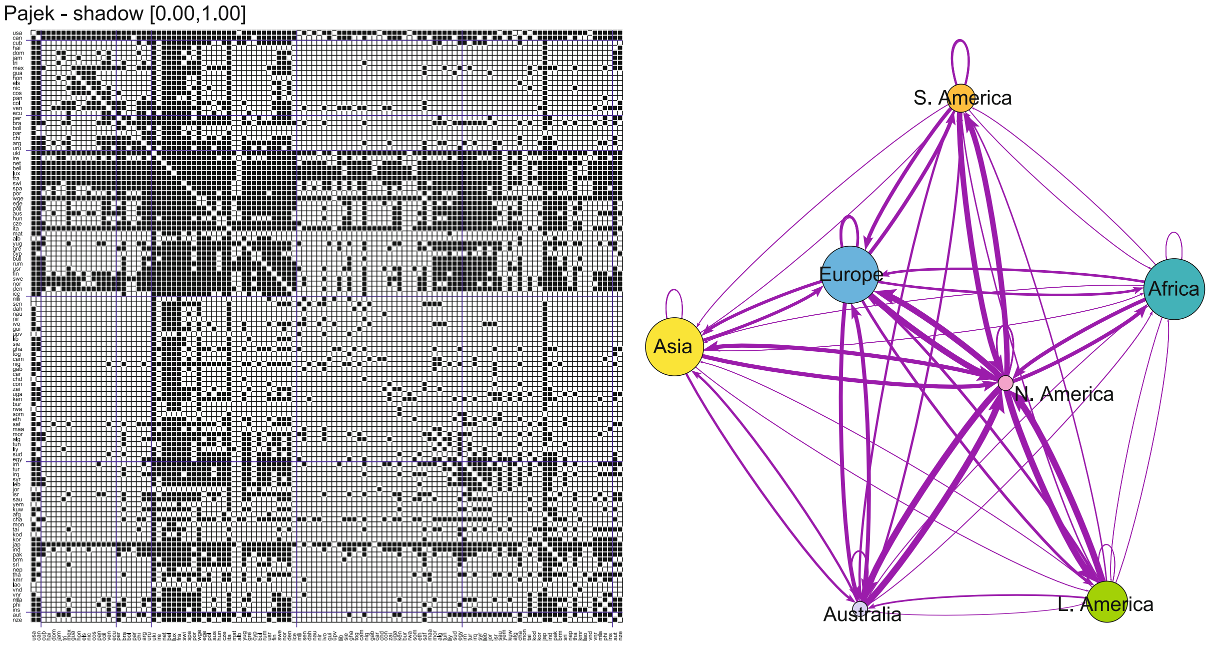
Contraction of cluster C in a graph $\mathcal{G}$ is called a graph $\mathcal{G}/C$, in which all vertices of the cluster C



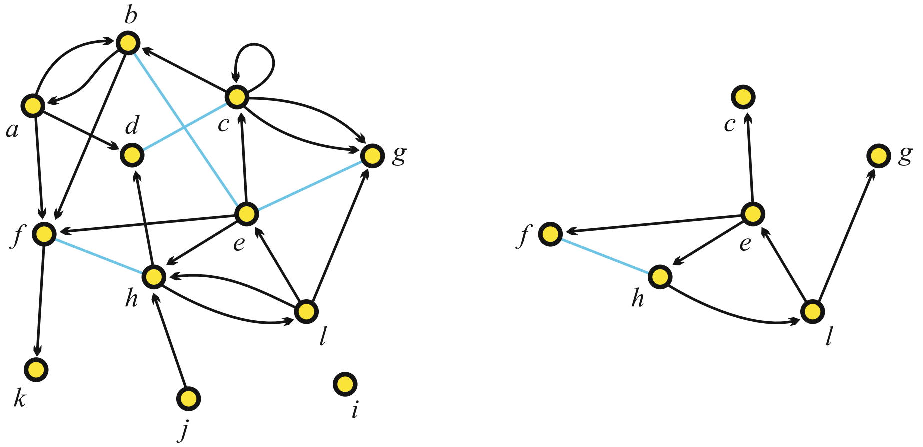
Social Network Analysis, Large-Scale, Figure 1
Decompositions

Social Network Analysis, Large-Scale, Table 2
Complexities of some typical algorithms

Algorithm	T(*n*)	1000	10 000	100 000	1 000 000	10 000 000
Alg-A	O(*n*)	0.00 s	0.015 s	0.17 s	2.22 s	22.2 s
Alg-B	O(*n* log *n*)	0.00 s	0.06 s	0.98 s	14.4 s	**2.8 m**
Alg-C	$O(n\sqrt{n})$	0.01 s	0.32 s	10.0 s	**5.27 m**	**2.78 h**
Alg-D	$O(n^2)$	0.07 s	7.50 s	**12.5 m**	**20.8 h**	**86.8 d**
Alg-E	$O(n^3)$	0.10 s	**1.67 m**	**1.16 d**	**3.17 y**	**3.17 ky**



Social Network Analysis, Large-Scale, Figure 2
Snyder and Kick's international trade; matrix display and reduction



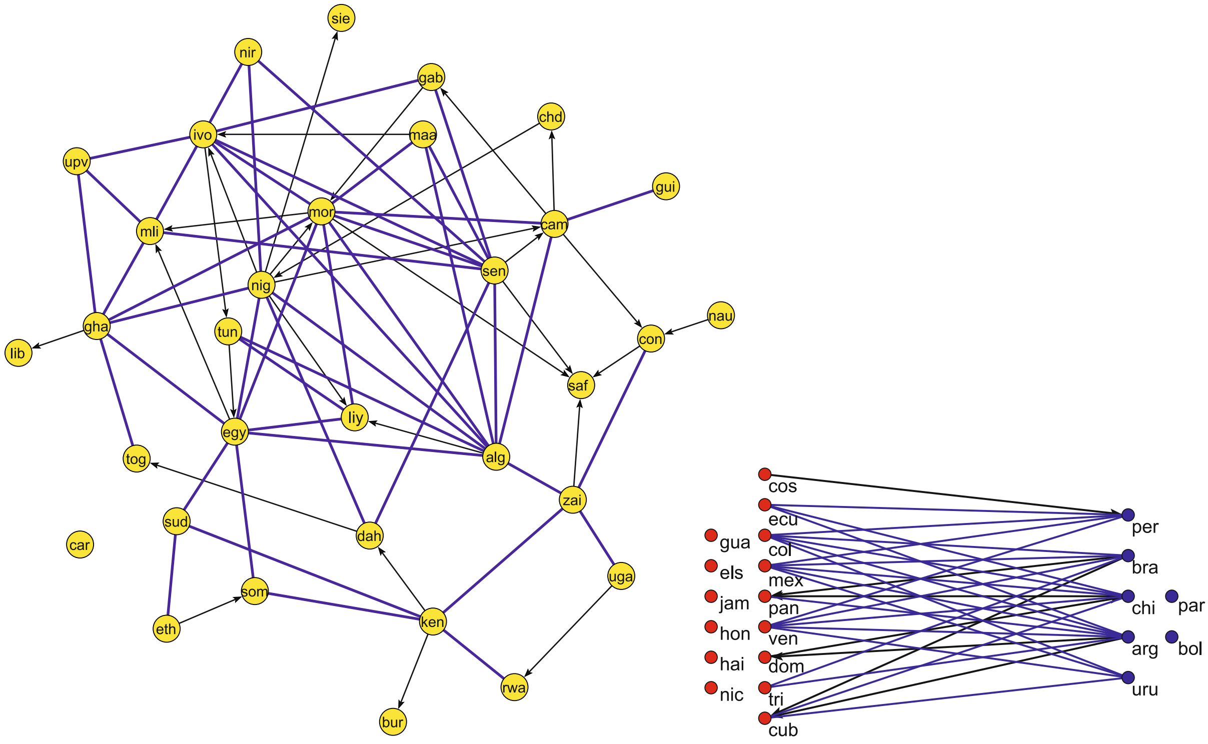
Social Network Analysis, Large-Scale, Figure 3
Graph and its subgraph

are replaced by a single new vertex, say c. More precisely: $\mathcal{G}/C = (\mathcal{V}', \mathcal{L}')$, where $\mathcal{V}' = (\mathcal{V} \setminus C) \cup \{c\}$ and $\mathcal{L}'$ consists of lines from $\mathcal{L}$ that have both end-points in $\mathcal{V} \setminus C$. Beside these it contains also a 'star' with the center c and: arc(v, c), if $\exists p \in \mathcal{L}, u \in C\colon\ p(v, u)$; or arc$(c, v)$, if $\exists p \in \mathcal{L}, u \in C\colon\ p(u, v)$. There is a loop (c, c) in c if $\exists p \in \mathcal{L}, u, v \in C\colon\ p(u, v)$.

In a network over graph $\mathcal{G}$, we have also to specify how the new values/weights are determined in the shrunk part of the network. Usually as the sum or maximum/minimum of the original values.

For a given partition if we contract all clusters except few selected we obtain their *context*; and if we contract all clusters we obtain the *reduction* of a given network.

On the left side of the Fig. 2 the *matrix display* of Snyder and Kick's [40] international trade network is presented. Vertices in the display are reordered according to the partition by (sub)continents. On the right side the



Social Network Analysis, Large-Scale, Figure 4
Africa cut-out and inter-links between South and Latin America

corresponding reduction of the network is presented. The lines in the reduction have the thickness proportional to the weights

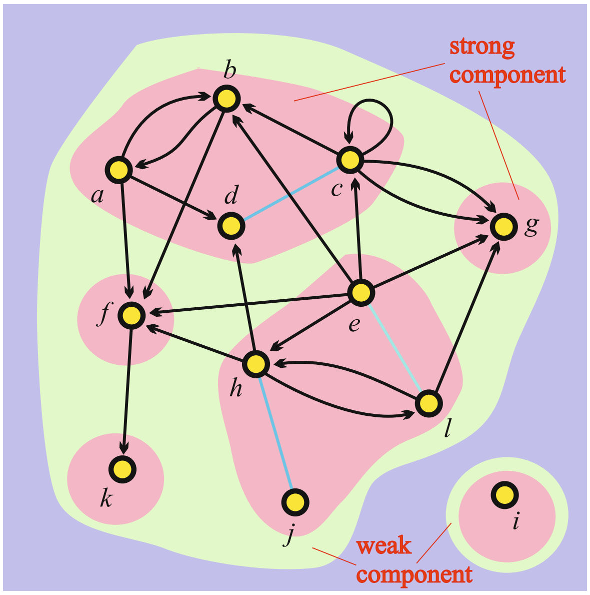
$$w(C_i, C_j) = \frac{n(C_i, C_j)}{n(C_i) \cdot n(C_j)}$$

where $n(C_i, C_j)$ is the number of lines from cluster C_i to cluster C_j; and $n(C_i)$ is the number of lines inside the cluster C_i.

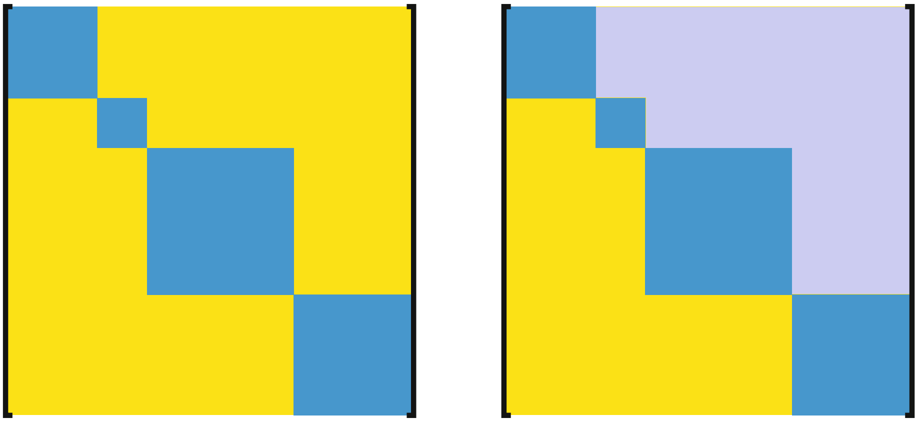
A *subgraph* $\mathcal{H} = (\mathcal{V}', \mathcal{L}')$ of a given graph $\mathcal{G} = (\mathcal{V}, \mathcal{L})$ is a graph which set of lines is a subset of set of lines of $\mathcal{G}$, $\mathcal{L}' \subseteq \mathcal{L}$, its vertex set is a subset of set of vertices of $\mathcal{G}$, $\mathcal{V}' \subseteq \mathcal{V}$, and it contains all end-vertices of $\mathcal{L}'$. The graph on the right side of Fig. 3 is a subgraph of the graph on the left side.

A subgraph can be *induced* by a given subset of vertices $\mathcal{V}'$, then $\mathcal{L}' = \mathcal{L}|\mathcal{V}'$ consists of all lines from $\mathcal{L}$ which have both end-points in $\mathcal{V}'$; or lines $\mathcal{L}'$, then $\mathcal{V}' = \mathcal{V}|\mathcal{L}'$ consists of all end-points of lines from $\mathcal{L}'$. It is a *spanning* subgraph iff $\mathcal{V}' = \mathcal{V}$.

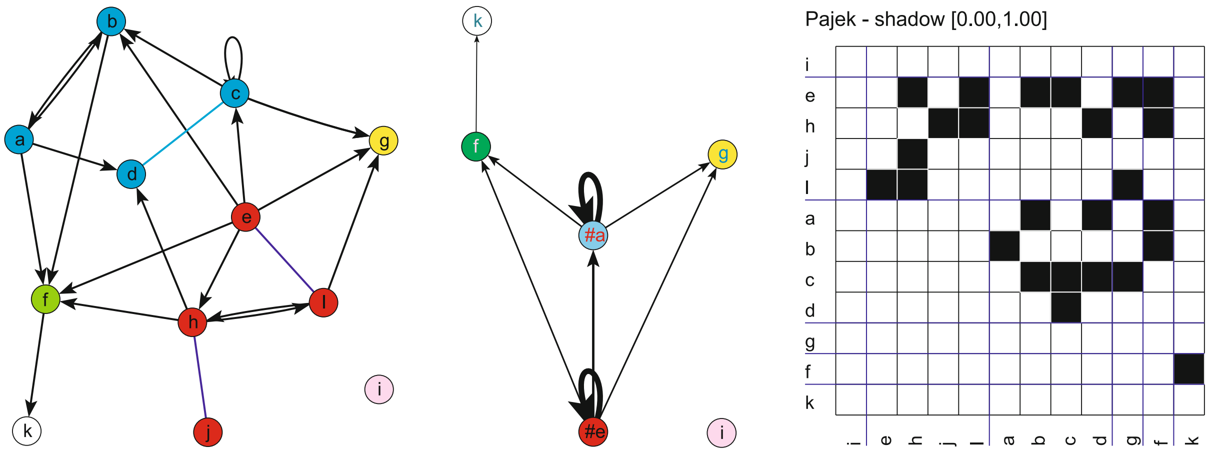
On the left side of Fig. 4 the *cut-out* of African countries from the Snyder and Kick's network is presented – the induced subgraph by Africa cluster; and on the right side the *inter-links* between Latin America and South America – the induced subgraph by Latin America and South America clusters with inside cluster lines removed.



Social Network Analysis, Large-Scale, Figure 5
Weak and strong components



Social Network Analysis, Large-Scale, Figure 6
Weak and strong components in matrix display



Pajek - shadow [0.00,1.00]

	i	e	h	j	l	a	b	c	d	g	f	k
i												
e			■		■		■	■		■	■	
h				■	■				■		■	
j			■									
l		■	■							■		
a							■		■		■	
b						■					■	
c							■	■	■	■		
d								■				
g												
f												■
k												

Social Network Analysis, Large-Scale, Figure 7
Condensation

Connectivity

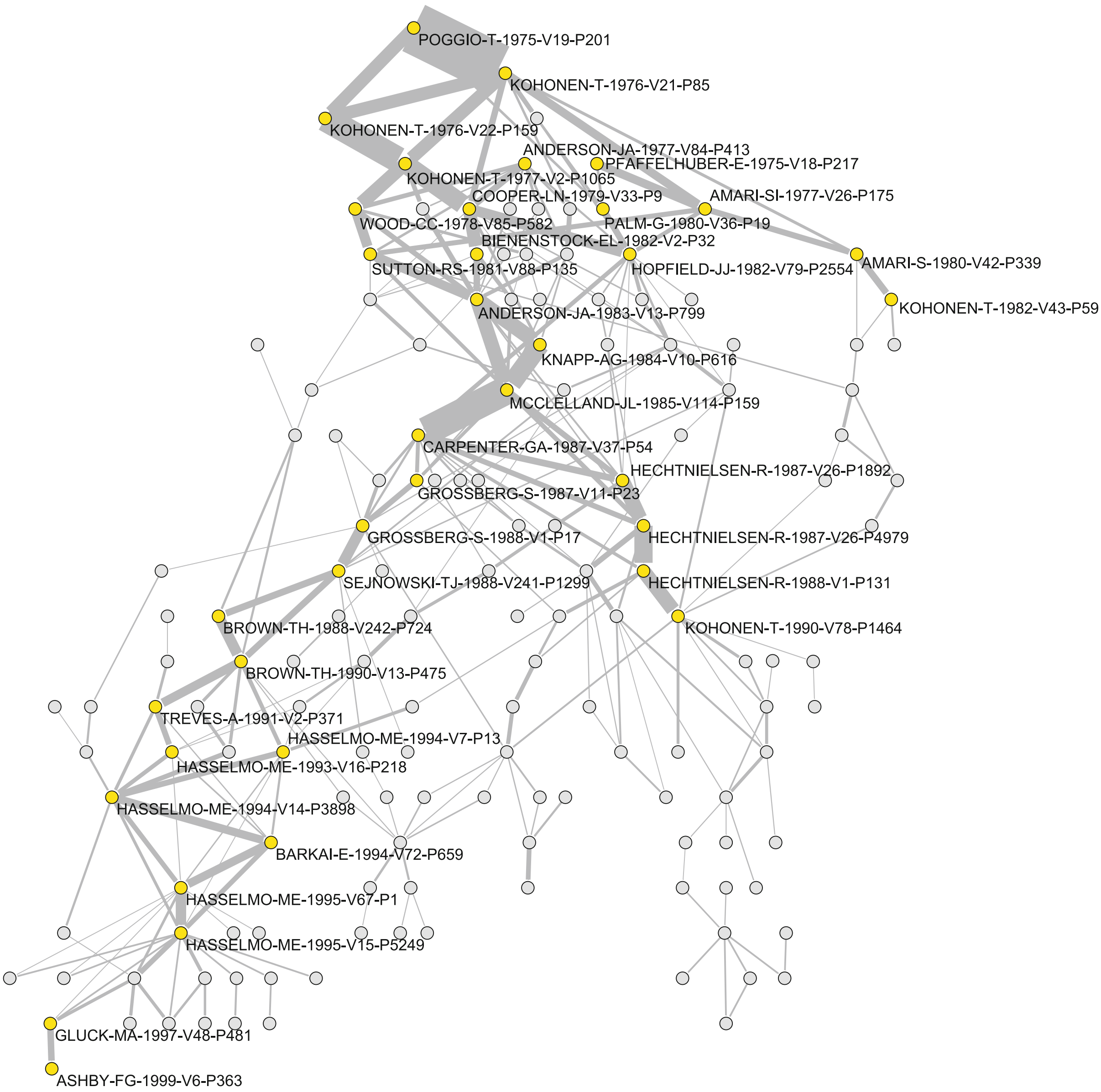
A *walk* from vertex u to vertex v is a sequence of lines $l(v_{i-1}, v_i)$, $i = 1, \dots, k$ such that $v_0 = u$ and $v_k = v$. k is called the *length* of the walk. If in the definition of a walk we don't care about the direction of its lines we get a *semiwalk*. A walk is *closed* iff $u = v$. A graph is *acyclic* iff it doesn't contain any closed walk. A walk in which all vertices are different is a *path*.

Vertex u is *reachable* from vertex v iff there exists a walk with initial vertex v and terminal vertex u. Vertex v is *weakly connected* with vertex u iff there exists a semiwalk with v and u as its end-vertices. Vertex v is *strongly connected* with vertex u iff they are mutually reachable.

Weak and strong connectivity are equivalence relations. Equivalence classes induce weak/strong *components* (See Fig. 5).

Reordering the vertices of network such that the vertices from the same class of weak partition are put together we get a matrix representation (left side of Fig. 6) consisting of diagonal blocks – weak components. The out-diagonal blocks are zero-blocks. Most problems can be solved separately on each component and afterward these solutions combined into final solution.

If we shrink every strong component of a given graph into a vertex, delete all loops and identify parallel arcs the obtained reduced graph, called also the *condensation* of a given graph, is acyclic [27]. For every acyclic



Social Network Analysis, Large-Scale, Figure 8
Main component of arc cut at level 0.007 of the SOM citation network

graph an *ordering/level* function $i \colon \mathcal{V} \to \mathbb{N}$ exists such that $(u, v) \in \mathcal{A} \Rightarrow i(u) < i(v)$. The procedure to determine such ordering is called *topological sort* [19]. Reordering in matrix display the vertices of a network by this ordering we obtain a representation as at the right side of Fig. 6 – the blocks below the diagonal are zero-blocks.

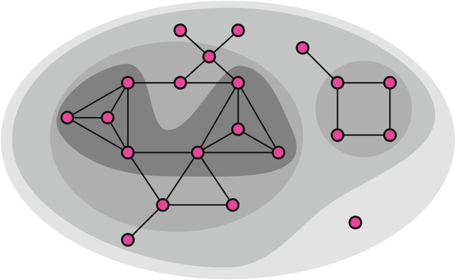
A directed graph, its condensation and its topologically ordered matrix display are presented in Fig. 7.

For several network analysis problems more efficient algorithms exist for acyclic networks.

Cuts

The basic approach to find interesting groups inside a network is to express our intentions (question) with an appropriate property/weight (measured or computed from network structure) and then identify the substructures of elements with the highest (lowest) values of the selected property. This approach is known as a method of *cuts*.

There exist several measures of importance of vertices in a network such as: degree, betweeness, closeness [16,24], hubs and authorities [30], clustering coefficient, …



Social Network Analysis, Large-Scale, Figure 9
Cores

The *degree* $\deg(v)$ of vertex v equals to the number of lines having vertex v as their end-point. The *maximum degree* of a graph is denoted by Δ. Similarly the in-degree $\operatorname{indeg}(v)$ of vertex v equals to the number of lines having vertex v as their terminal point, and the out-degree $\operatorname{outdeg}(v)\ldots$

The *vertex-cut* of a network $\mathcal{N} = (\mathcal{V}, \mathcal{L}, p)$, for a property $p\colon \mathcal{V} \to \mathbb{R}$, at selected level t is a subnetwork $\mathcal{N}(t) = (\mathcal{V}', \mathcal{L}(\mathcal{V}'), p)$, determined by the set

$$\mathcal{V}' = \{v \in \mathcal{V}\colon p(v) \geq t\}$$

and $\mathcal{L}(\mathcal{V}')$ is the set of lines from $\mathcal{L}$ that have both end-points in $\mathcal{V}'$.

The *line-cut* of a network $\mathcal{N} = (\mathcal{V}, \mathcal{L}, w)$, for a weight $w\colon \mathcal{L} \to \mathbb{R}$, at selected level t is a subnetwork $\mathcal{N}(t) = (\mathcal{V}(\mathcal{L}'), \mathcal{L}', w)$, determined by the set

$$\mathcal{L}' = \{e \in \mathcal{L}\colon w(e) \geq t\}$$

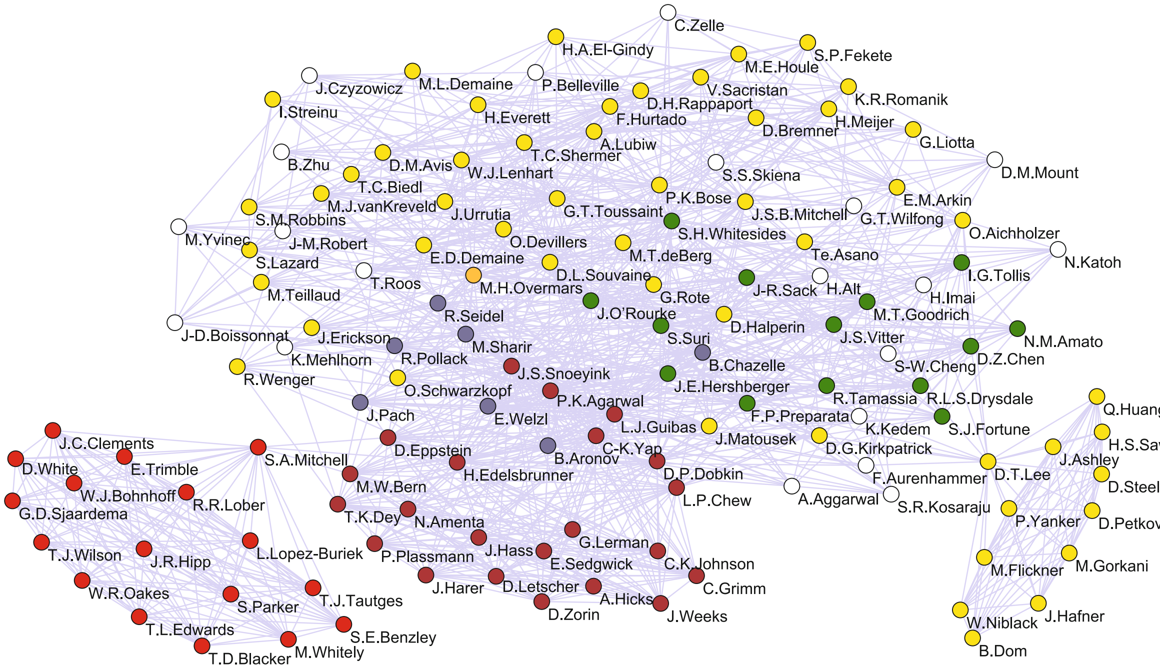
and $\mathcal{V}(\mathcal{L}')$ is the set of all end-points of the lines from $\mathcal{L}'$.

In the analysis of a cut $\mathcal{N}(t)$, we look at its components. Their number and sizes depend on t. Usually there are many small components. Often we consider only components of size at least k and not exceeding K. The components of size smaller than k are discarded as 'less interesting'; and the components of size larger than K are cut again at some higher level.

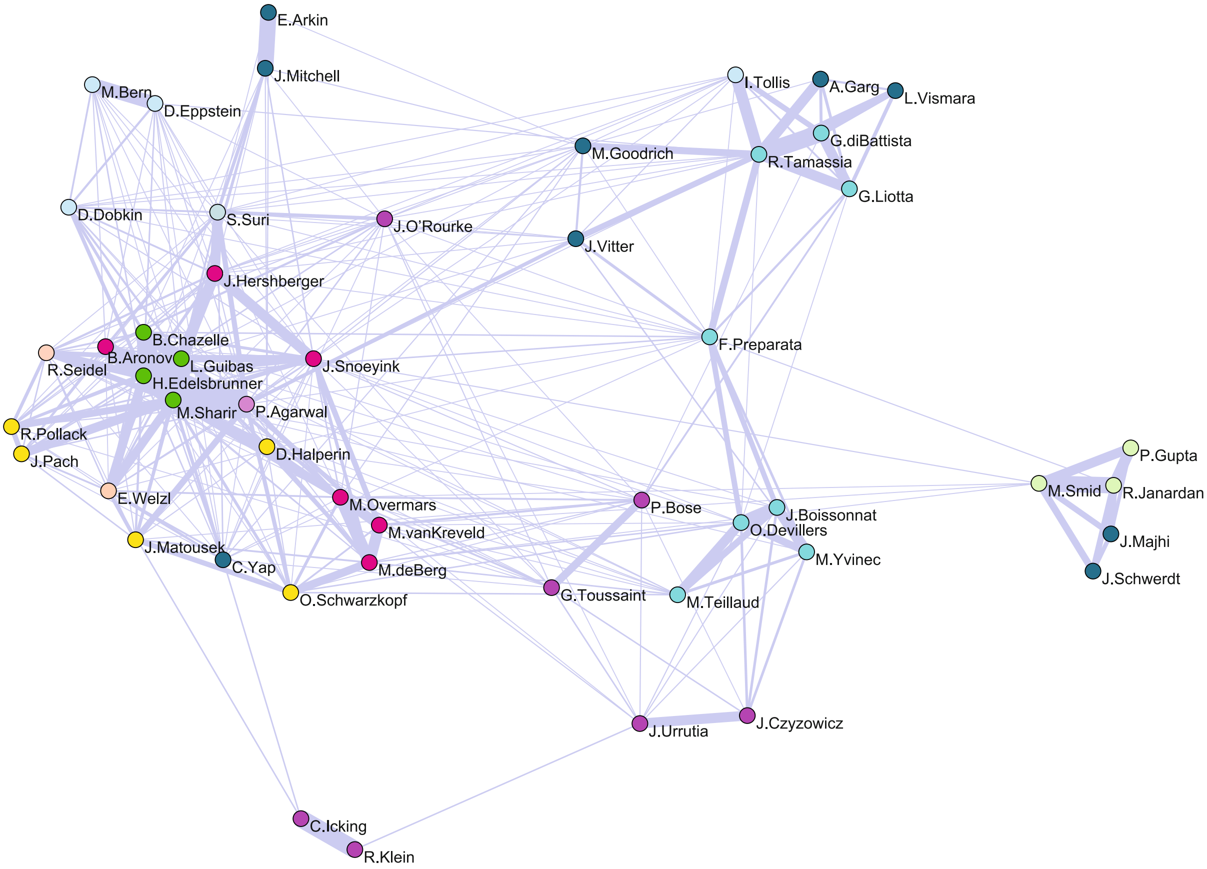
The values of threshold t and size bounds k and K are determined by inspecting the distribution of vertex/line-values and the distribution of component sizes and considering additional knowledge on the nature of network or goals of analysis.

The p_S-core at level 46 (see Fig. 11) of the collaboration network in the field of computational geometry is an example of vertex cut.

The citation network analysis started in 1964 with the paper of Garfield et al. [25]. In 1989 Hummon and Doreian [29] proposed three indices – weights of arcs that are proportional to the number of different source-sink paths passing through the arc. In Fig. 8 the main component



Social Network Analysis, Large-Scale, Figure 10
Cores of orders 10–21 in Computational Geometry collaboration network



Social Network Analysis, Large-Scale, Figure 11
p_S-core at level 46 in Computational Geometry collaboration network

of the arc cut at level 0.007 for SPC (search path count) weights of the SOM (selforganizing maps) citation network (4470 vertices, 12,731 arcs) is presented.

Dense Groups – Cores and Short Rings

Several notions were proposed in attempts to formally describe dense groups in graphs.

Clique of order k, $k \geq 3$, is a maximal complete subgraph (isomorphic to complete graph K_k – graph with k vertices and all possible edges among them).

Other notions are: s-plexes, s-clans, LS sets, lambda sets, cores, … (Wasserman and Faust [43]). For all of them, except for cores, it turned out that they are difficult (no fast algorithm exists) to determine.

The notion of core was introduced by Seidman in 1983 [38]. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph. A subgraph $\mathcal{H}_k = (\mathcal{W}, \mathcal{E}|\mathcal{W})$ induced by the set $\mathcal{W}$ is a *k-core* or a *core of order k* iff for all $v \in \mathcal{W}\colon \deg_{\mathcal{H}_k}(v) \geq k$, and $\mathcal{H}_k$ is a maximal subgraph with this property. The core of maximum order is also called the *main* core. The *core number* of vertex v is the highest order of a core that contains this vertex. In general graphs instead of the degree $\deg(v)$ we can also use: in-degree, out-degree, in-degree + out-degree, etc., determining different types of cores.

From Fig. 9, representing 0, 1, 2 and 3 core, we can see the following properties of cores:

- The cores are nested: $i < j \Longrightarrow \mathcal{H}_j \subseteq \mathcal{H}_i$. They form a hierarchy.
- Cores are not necessarily connected subgraphs.

An efficient algorithm for determining the cores hierarchy is based on the following property: If from a given graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ we recursively delete all vertices, and edges incident with them, of degree less than k, the remaining graph is the k-core.

The Fig. 10 presents the cores of orders 10 to 21 in the collaboration network ($n = 7343$, $m = 11{,}898$) for the field of computational geometry – two authors are linked iff they wrote a paper together. The weight of the edge equals to the number of joint papers.

The notion of core can be generalized to networks. Let $\mathcal{N} = (\mathcal{V}, \mathcal{E}, w)$ be a network, where $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a graph and weight $w\colon \mathcal{E} \to \mathbb{R}$ is a function assigning values to edges. A *vertex property function* on $\mathcal{N}$, or a *p-function* for short, is a function $p(v, U), v \in \mathcal{V}, U \subseteq \mathcal{V}$ with real values. Let $N_U(v) = N(v) \cap U$, where $N(v)$ is the set of neighbors of v. Besides degrees, here are some other examples of p-functions [13]:

$$
\begin{aligned}
p_S(v,U) &= \sum_{u \in N_U(v)} w(v,u), \text{ where } w\colon \mathcal{E} \to \mathbb{R}_0^+ \\
p_M(v,U) &= \max_{u \in N_U(v)} w(v,u), \text{ where } w\colon \mathcal{E} \to \mathbb{R} \\
p_k(v,U) &= \text{number of cycles of length } k \\
&\qquad \text{through vertex } v \text{ in } (U, \mathcal{E}|U) \\
p_\gamma(v,U) &= \frac{\deg(v,U)}{\max_{u \in N(v)} \deg(u)}, \text{ if } \deg(v) > 0; \\
&\qquad\qquad\qquad 0, \text{ otherwise} \\
p_\delta(v,U) &= \max_{u \in N_U^+(v)} \deg(u) - \min_{u \in N_U^+(v)} \deg(u) \\
p_a(v,U) &= \frac{1}{|N_U(v)|} \sum_{u \in N_U(v)} w(v,u), \text{ if } N_U(v) \neq \emptyset; \\
&\qquad\qquad\qquad 0, \text{ otherwise}
\end{aligned}
$$

The subgraph $\mathcal{H} = (C, \mathcal{E}|C)$ induced by the set $C \subseteq \mathcal{V}$ is a *p-core at level* p-core $t \in \mathbb{R}$ iff for all $v \in C\colon t \leq p(v, C)$ and C is a maximal such set.

The function p is *monotone*, iff it has the property

$$C_1 \subset C_2 \Rightarrow \forall v \in \mathcal{V}\colon (p(v, C_1) \leq p(v, C_2))$$

The degrees and the functions p_S, p_M, p_k, p_γ and p_δ are monotone; and p_a is not. For a monotone function the p-core at level t can be determined, as in the ordinary case, by successively deleting vertices with value of p lower than t; and the cores on different levels are nested

$$t_1 < t_2 \Rightarrow \mathcal{H}_{t_2} \subseteq \mathcal{H}_{t_1}$$

The p-function is *local*, iff $p(v, U) = p(v, N_U(v))$. The degrees, $p_S, p_M, p_\gamma, p_\delta$ and p_a are local; but p_k is **not** local for $k \geq 4$. For a local monotone p-function an $O(m \max(\Delta, \log n))$ algorithm for determining the p-core levels exists, assuming that $p(v, N_C(v))$ can be computed in $O(\deg_C(v))$.

Figure 11 presents the p_S-core at level 46 of the collaboration network in the field of computational geometry. Note, for example, that R. Klein (lower left) has in-core degree only 2, but its in-core sum of weights is at least 46 – he wrote most of his papers with C. Icking.

A k-ring is a simple closed chain of length k. Using k-rings we can define a weight of an edge e as $w_k(e) = \#$ of different k-rings containing the edge $e \in \mathcal{E}$.

Since for each edge e of complete graph K_r, $r \geq k \geq 3$ we have $w_k(e) = (r-2)!/(r-k)!$, the edges belonging to cliques have large weights. Therefore these weights can be used to identify the dense parts of a network. For example: all r-cliques of a network belong to $(r-2)$-edge cut for the weight w_3.

Related to triangular (3-rings) network is the notion of *triangular connectivity* that can be used to operationalize the notion of Granovetter's strong and weak ties [26]. This notion can be generalized to short cycle connectivity. For details see [14]. For efficient algorithms for computing triangles in networks see [10,39], and Latapy.

In Fig. 12 the edge-cut at level 16 of triangular network of Erdős collaboration graph (without Erdős, $n = 6926$, $m = 11{,}343$) is presented [9].

In directed networks there are two types of triangles or 3-rings (cyclic and transitive, see Fig. 14).

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a simple undirected graph. *Clustering* in vertex v is usually measured as a quotient between the number of lines in subgraph $\mathcal{G}^1(v) = \mathcal{G}(N(v))$ induced by the neighbors of vertex v and the number of lines in the complete graph on these vertices:

$$C(v) = \begin{cases} \dfrac{2|\mathcal{L}(\mathcal{G}^1(v))|}{\deg(v)(\deg(v)-1)}, & \deg(v) > 1 \\ 0, & \text{otherwise} \end{cases} .$$

For simple directed graphs, we have to omit the number 2.

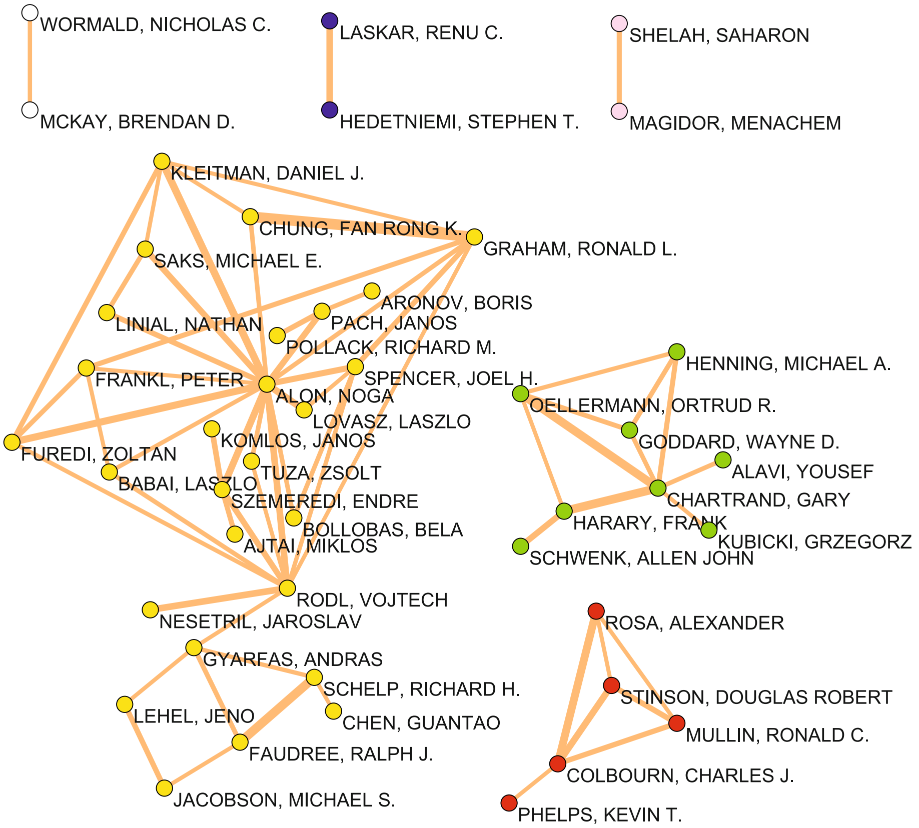
So defined clustering coefficient attains largest values mostly on vertices of low degree – it is not useful for data analysis task. A better coefficient is obtained by the following correction

$$C_1(v) = \frac{\deg(v)}{\Delta} C(v)$$

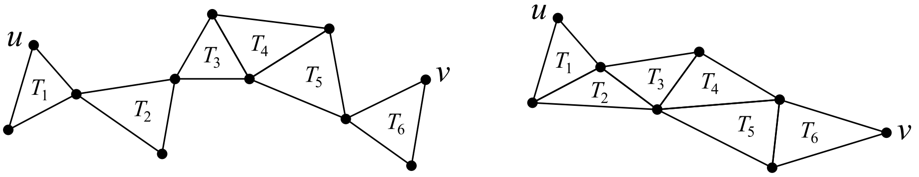
where Δ is the maximum degree in graph $\mathcal{G}$. This measure attains its largest value in vertices that belong to an isolated clique of size Δ.

Islands

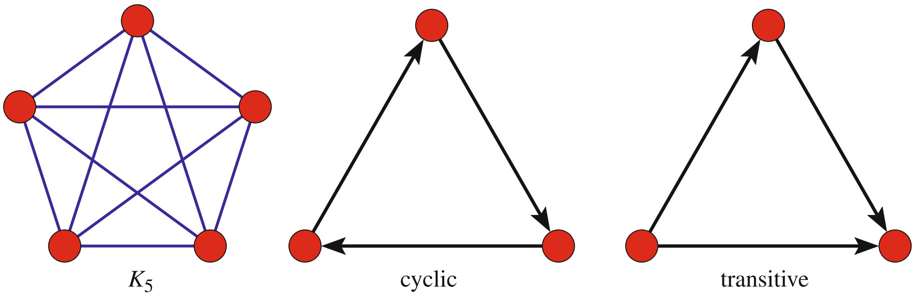
Islands are very general and efficient approach to determine the 'important' subnetworks in a given network with respect to a given property of vertices or lines. It is an improvement of the cuts approach. If we represent a given or computed value of vertices/lines as a height of vertices/



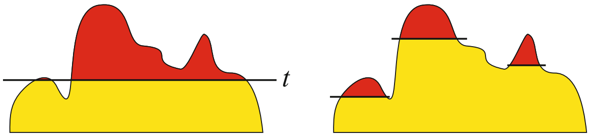
Social Network Analysis, Large-Scale, Figure 12
Edge-cut at level 16 of triangular network of Erdős collaboration graph



Social Network Analysis, Large-Scale, Figure 13
Vertex and edge triangular connectivity



Social Network Analysis, Large-Scale, Figure 14
K_5 and cyclic and transitive 3-ring

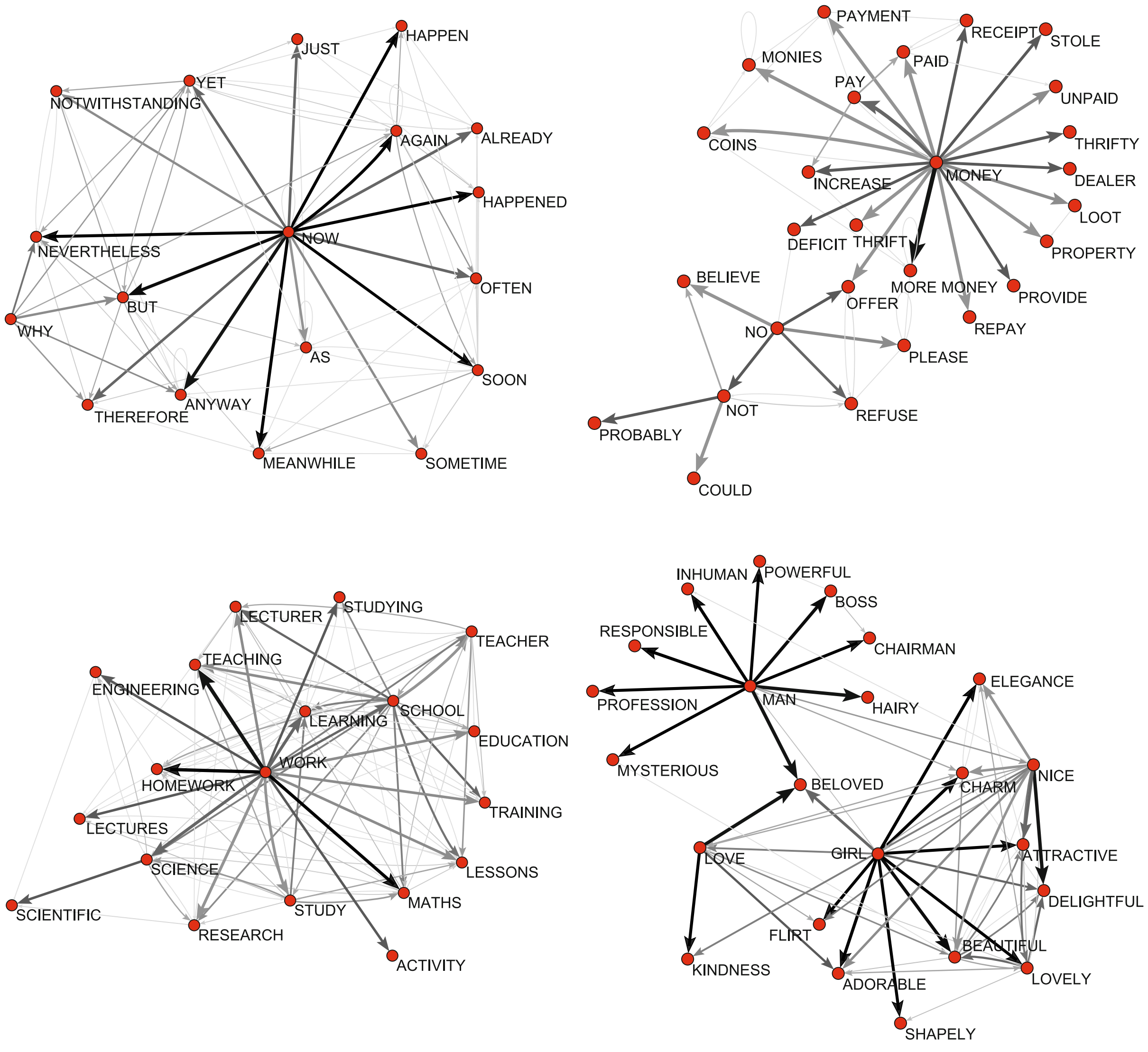


Social Network Analysis, Large-Scale, Figure 15
Cuts and islands

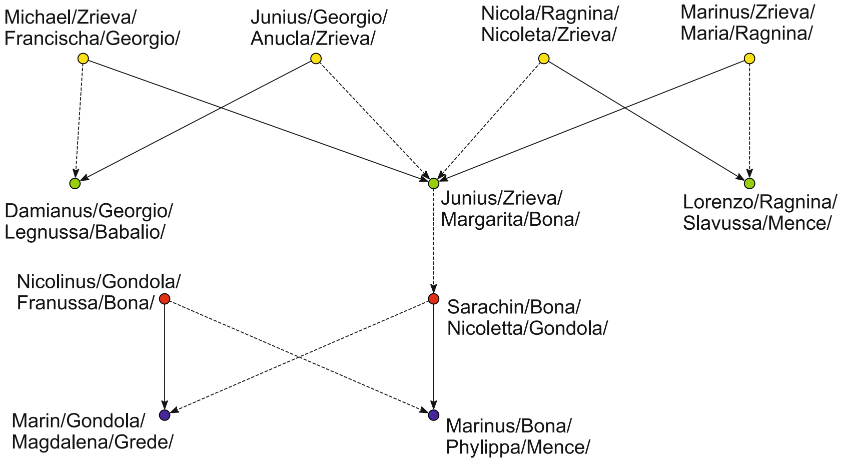
lines and we immerse the network into a water up to selected level we get *islands*. Varying the level we get different islands [45].

In the islands approach we select only maximal islands of sizes inside the given size bounds k to K, but on different levels. In this way we bypass the problems of the cuts approach: determining the 'right' threshold value and too small/large sizes of obtained components. Besides this we can also identify locally important islands with small heights – emerging groups. Very efficient algorithms exist to determine the islands hierarchy and to list all the islands of selected sizes. An island is *simple* iff it has only one peak.

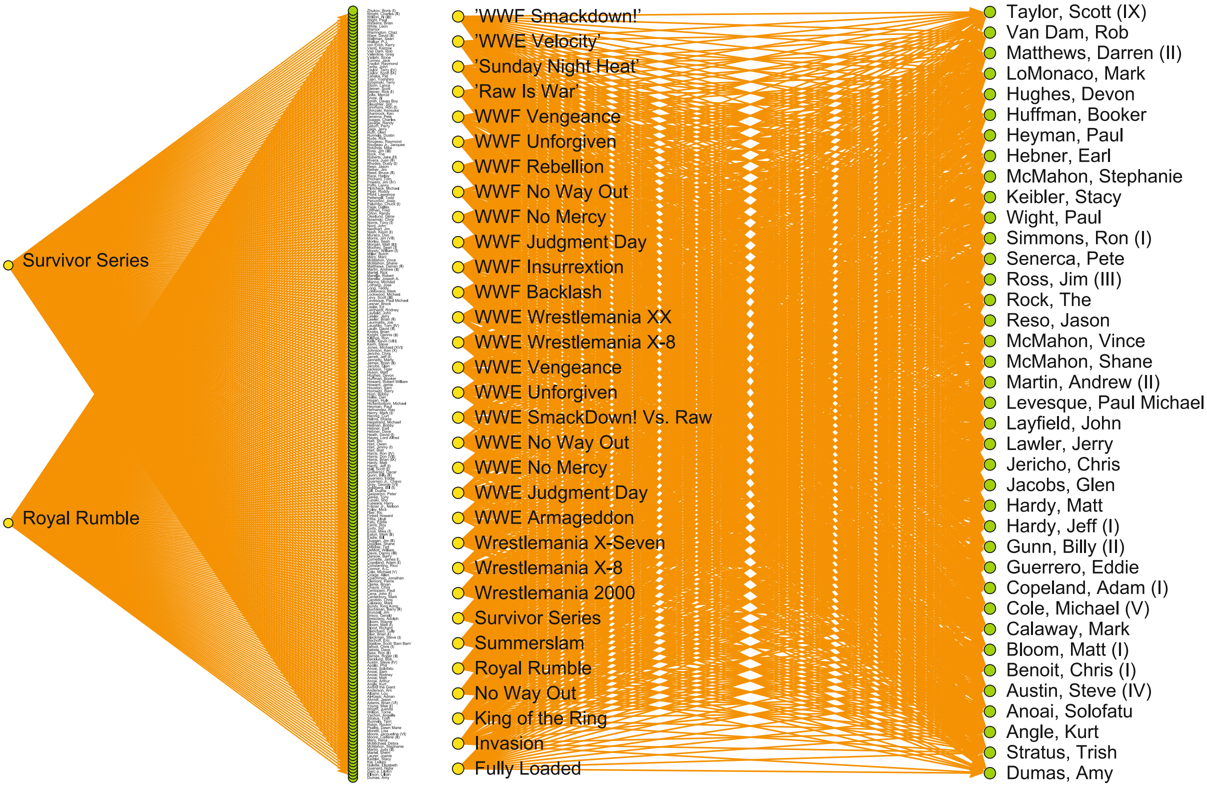
As an example, let us take the Nber network of US Patents. It has 3,774,768 vertices and 16,522,438 arcs. We computed SPC weights in it and determined all (2,90)-islands. The reduced network has 470,137 vertices, 307,472 arcs and for different k: $C_2 = 187{,}610$, $C_5 = 8859$, $C_{30} = 101$, $C_{50} = 30$ islands. The main island turns out



Social Network Analysis, Large-Scale, Figure 16
Selected islands from The Edinburgh Associative Thesaurus



Social Network Analysis, Large-Scale, Figure 17
Marriages among relatives in Ragusa

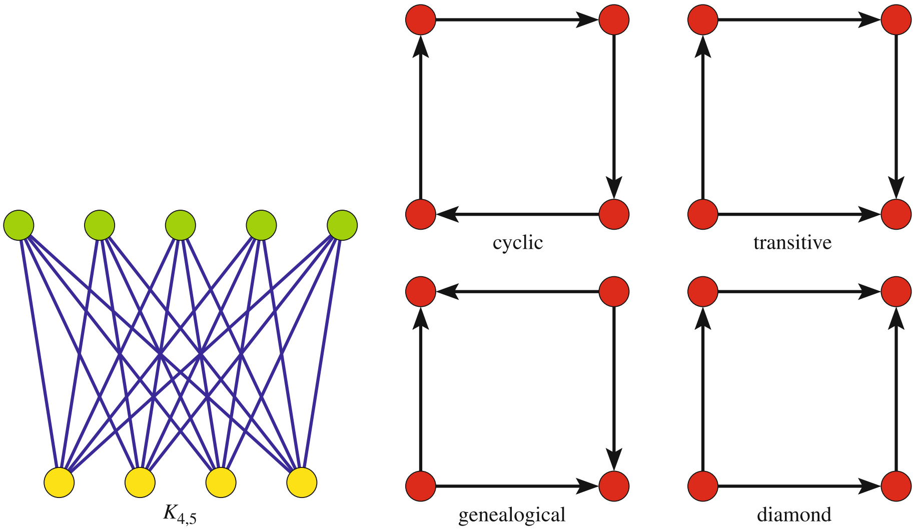


Social Network Analysis, Large-Scale, Figure 18
(247,2)-core and (27,22)-core of IMDB – wrestling

to be the island on the theme *LCD – Liquid crystal display*. In Fig. 16 four islands for transitivity triangular weight from *The Edinburgh Associative Thesaurus* network ($n = 23{,}219$, $m = 325{,}624$) are presented. From the left bottom island of words around the leader 'WORK' we see that the data were collected asking students.

Pattern Searching

If a selected *pattern* determined by a given graph does not occur frequently in a sparse network the straightforward backtracking algorithm applied for pattern searching finds all appearances of the pattern very fast even in



Social Network Analysis, Large-Scale, Figure 19
$K_{4,5}$ and directed 4-rings

the case of very large networks. Pattern searching was successfully applied to searching for patterns of atoms in large organic molecula (carbon rings) and searching for relinking marriages in genealogies (Batagelj and Mrvar [12]; Batagelj [6]).

The Fig. 17 presents three connected relinking marriages in the genealogy (represented as a p-graph) of Ragusan noble families. In a p-graph the vertices represent married couples or nonmarried individuals. A solid arc indicates the _ *is a son of* _ relation, and a dotted arc indicates the _ *is a daughter of* _ relation. In all three patterns a brother and a sister from one family found their partners in the same other family.

Two Mode Networks

A network $\mathcal{N} = (\mathcal{V}, \mathcal{L}, w)$ in which the set of vertices $\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2$ is composed of two disjoint sets $\mathcal{V}_1$ and $\mathcal{V}_2$, and $\mathcal{L}$ is a set of *lines* linking $\mathcal{V}_1$ and $\mathcal{V}_2$ is called a *two-mode* or bipartite network.

The two-mode networks often appear in applications, but till recently no directed methods for analysis of larger two-mode networks were available. To identify dense parts of two-mode network, we can use the adapted cores and short rings approaches (Ahmed et al. [3]).

The subset of vertices $C \subseteq \mathcal{V}$ is a (p, q)-core in a two-mode network $\mathcal{N} = (\mathcal{V}_1, \mathcal{V}_2; \mathcal{L})$, $\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2$ iff

a. in the induced subnetwork $\mathcal{H} = (C_1, C_2; \mathcal{L}(C))$, $C_1 = C \cap \mathcal{V}_1$, $C_2 = C \cap \mathcal{V}_2$ it holds for all $v \in C_1$: $\deg_{\mathcal{H}}(v) \geq p$ and for all $v \in C_2$: $\deg_{\mathcal{H}}(v) \geq q$;

b. C is the maximal subset of $\mathcal{V}$ satisfying condition **a**.

The two-mode cores have the following properties:

- $C(0, 0) = \mathcal{V}$
- $C(p, q)$ is not always connected
- $(p_1 \leq p_2) \wedge (q_1 \leq q_2) \Rightarrow C(p_2, q_2) \subseteq C(p_1, q_1)$.

To determine a (p, q)-core an algorithm similar to the ordinary core algorithm can be used: recursively remove from the first set all vertices of degree less than p, and from the second set all vertices of degree less than q. It can be implemented to run in $O(m)$ time.

The main question when applying the bipartite cores is what are the right values of p and q? The most interesting are the values on the 'border' that don't produce too large cores.

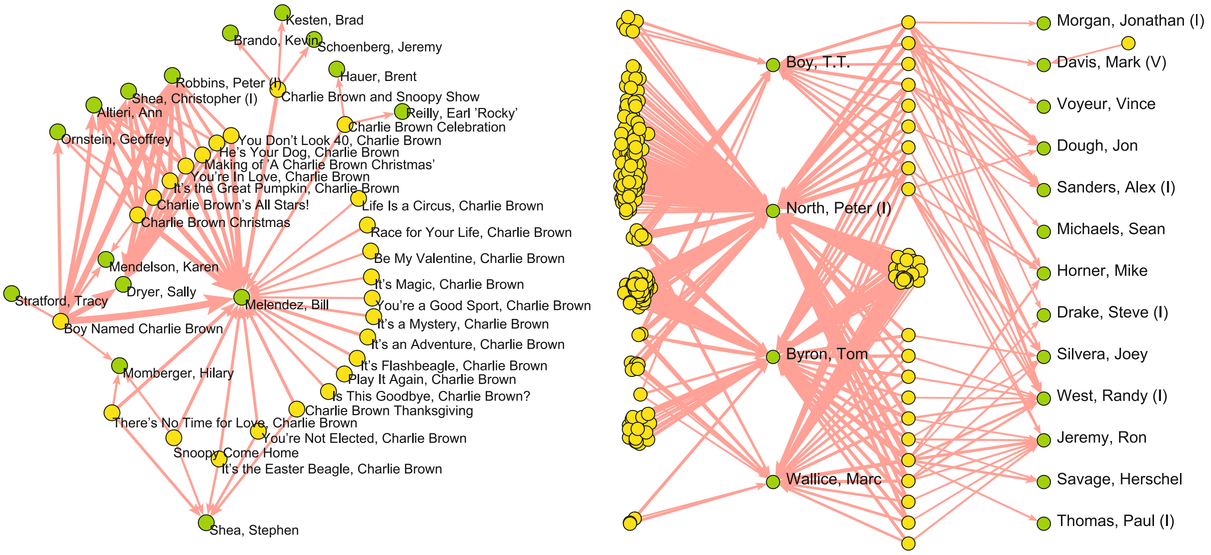
In Fig. 18 the (247,2)-core and (27,22)-core from the Internet Movie Database (two-mode network actors × movies, $n = 1{,}324{,}748 = 428{,}440 + 896{,}308$ vertices and $m = 3{,}792{,}390$ arcs) are presented. Both deal with wrestling.

In 2-mode network there are no 3-rings. The densest substructures are complete bipartite subgraphs $K_{p,q}$ – see $K_{4,5}$ on the left side of Fig. 19. They contain many 4-rings

$$w_4(e) = (p-1)(q-1)\,, \quad \text{for } \; e \in K_{p,q}\,.$$

There are four types of directed 4-rings – see the right side of Fig. 19. In the case of transitive rings we can count also on how many transitive rings the arc is a *shortcut*.

In the Internet Movie Database, we obtained for w_4 12,465 simple line islands on 56,086 vertices; 30 among


Kesten, Brad
Brando, Kevin
Schoenberg, Jeremy
Hauer, Brent
Robbins, Peter (I)
Charlie Brown and Snoopy Show
Shea, Christopher (I)
Reilly, Earl 'Rocky'
Altieri, Ann
Charlie Brown Celebration
Ornstein, Geoffrey
You Don't Look 40, Charlie Brown
He's Your Dog, Charlie Brown
Making of 'A Charlie Brown Christmas'
You're In Love, Charlie Brown
It's the Great Pumpkin, Charlie Brown
Charlie Brown's All Stars!
Life Is a Circus, Charlie Brown
Charlie Brown Christmas
Race for Your Life, Charlie Brown
Be My Valentine, Charlie Brown
Mendelson, Karen
It's Magic, Charlie Brown
Dryer, Sally
Stratford, Tracy
Melendez, Bill
You're a Good Sport, Charlie Brown
It's a Mystery, Charlie Brown
Boy Named Charlie Brown
It's an Adventure, Charlie Brown
It's Flashbeagle, Charlie Brown
Momberger, Hilary
Play It Again, Charlie Brown
Is This Goodbye, Charlie Brown?
Charlie Brown Thanksgiving
There's No Time for Love, Charlie Brown
You're Not Elected, Charlie Brown
Snoopy Come Home
It's the Easter Beagle, Charlie Brown
Shea, Stephen
Morgan, Jonathan (I)
Boy, T.T.
Davis, Mark (V)
Voyeur, Vince
Dough, Jon
Sanders, Alex (I)
North, Peter (I)
Michaels, Sean
Horner, Mike
Drake, Steve (I)
Byron, Tom
Silvera, Joey
West, Randy (I)
Jeremy, Ron
Wallice, Marc
Savage, Herschel
Thomas, Paul (I)

Social Network Analysis, Large-Scale, Figure 20
Islands for w_4 / Charlie Brown and adult

them have size at least 50. Two of them are presented on Fig. 20.

Multiplication of Networks

To a simple two-mode *network* $\mathcal{N} = (\mathcal{I}, \mathcal{J}, \mathcal{E}, w)$; where $\mathcal{I}$ and $\mathcal{J}$ are sets of *vertices*, $\mathcal{E}$ is a set of *edges* linking $\mathcal{I}$ and $\mathcal{J}$, and $w\colon \mathcal{E} \to \mathbb{R}$ is a *weight*; we can assign a *network matrix* $\mathbf{W} = [w_{i,j}]$ with elements: $w_{i,j} = w(i,j)$ for $(i,j) \in \mathcal{E}$ and $w_{i,j} = 0$ otherwise.

Given a pair of compatible networks $\mathcal{N}_A = (\mathcal{I}, \mathcal{K}, \mathcal{E}_A, w_A)$ and $\mathcal{N}_B = (\mathcal{K}, \mathcal{J}, \mathcal{E}_B, w_B)$ with corresponding matrices $\mathbf{A}_{\mathcal{I}\times\mathcal{K}}$ and $\mathbf{B}_{\mathcal{K}\times\mathcal{J}}$ we call a *product of networks* $\mathcal{N}_A$ and $\mathcal{N}_B$ a network $\mathcal{N}_A \star \mathcal{N}_B = \mathcal{N}_C = (\mathcal{I}, \mathcal{J}, \mathcal{E}_C, w_C)$, where $\mathcal{E}_C = \{(i,j)\colon i \in \mathcal{I}, j \in \mathcal{J}, c_{i,j} \neq 0\}$ and $w_C(i,j) = c_{i,j}$ for $(i,j) \in \mathcal{E}_C$. The product matrix $\mathbf{C} = [c_{i,j}]_{\mathcal{I}\times\mathcal{J}} = \mathbf{AB}$ is defined in the standard way

$$c_{i,j} = \sum_{k\in\mathcal{K}} a_{i,k} \cdot b_{k,j} \,.$$

In the case when $\mathcal{I} = \mathcal{K} = \mathcal{J}$ we are dealing with ordinary one-mode networks (with square matrices).

The standard matrix multiplication is too slow to be used for large networks. For sparse large networks, we can multiply much faster considering only nonzero elements. In general the multiplication of large sparse networks is a 'dangerous' operation since the result can 'explode' – it is not sparse. But in many interesting cases, we can assure that also the product is sparse. For example, we can prove:

If at least one of the sparse networks $\mathcal{N}_A$ and $\mathcal{N}_B$ has small maximal degree on K then also the resulting product network $\mathcal{N}_C$ is sparse.

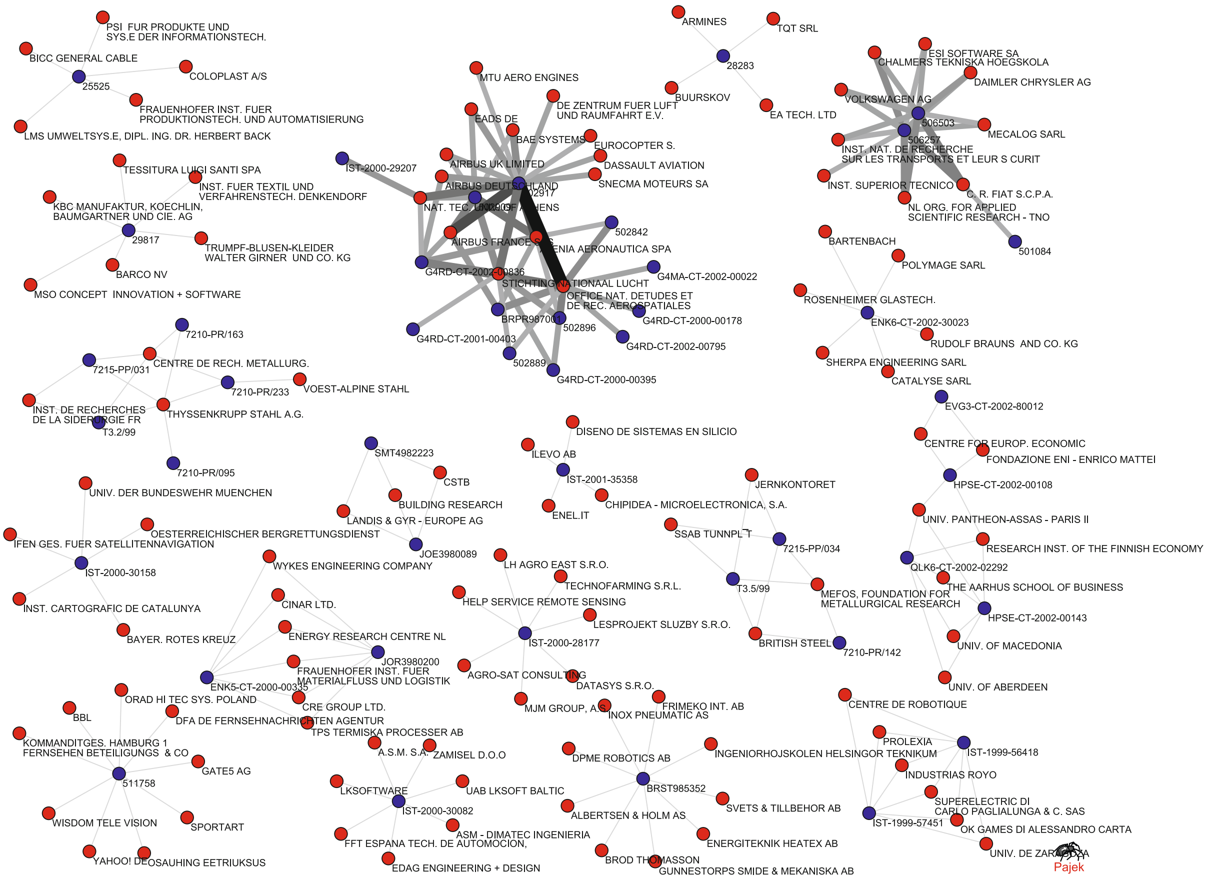
A more detailed analysis gives: Let for $k \in \mathcal{K}$ be $d_{\min}(k) = \min(\deg_A(k), \deg_B(k))$, $\Delta_{\min} = \max_{k\in\mathcal{K}} d_{\min}(k)$, $d_{\max}(k) = \max(\deg_A(k), \deg_B(k))$, $\mathcal{K}(d) = \{k \in \mathcal{K}\colon d_{\max}(k) \geq d\}$, and $d^* = \operatorname{argmin}_d (|\mathcal{K}(d)| \leq d)$. If for the sparse networks $\mathcal{N}_A$ and $\mathcal{N}_B$ the quantities $\Delta_{\min}$ and d^* are small then also the resulting product network $\mathcal{N}_C$ is sparse.

For example, using network multiplication we can in a given genealogy from the basic relations (P – parent-of, L – is a man, J – is a woman) compute all other kinship relations. For details see [12].

An important application of network multiplication is conversion of two-mode network to the corresponding one-mode networks. Often we transform a two-mode network $\mathcal{N}$ into an ordinary (one-mode) network $\mathcal{N}_1 = (\mathcal{I}, \mathcal{E}_1, w_1)$ or/and $\mathcal{N}_2 = (\mathcal{J}, \mathcal{E}_2, w_2)$, where $\mathcal{E}_1$ and w_1 are determined by the matrix $\mathbf{W}^{(1)} = \mathbf{W}\mathbf{W}^{\mathrm{T}}$, $w^{(1)}_{ij} = \sum_{k\in\mathcal{J}} w_{ik} \cdot w^{\mathrm{T}}_{kj}$ and $\mathbf{W}^{\mathrm{T}}$ is the transpose of matrix $\mathbf{W}$. Evidently the matrix $\mathbf{W}^{(1)}$ is symmetric $w^{(1)}_{ij} = w^{(1)}_{ji}$. There is an edge $(i,j) \in \mathcal{E}_1$ in $\mathcal{N}_1$, iff $N(i) \cap N(j) \neq \emptyset$. Its weight is $w_1(i,j) = w^{(1)}_{ij}$. The network $\mathcal{N}_2$ is determined in a similar way by the matrix $\mathbf{W}^{(2)} = \mathbf{W}^{\mathrm{T}}\mathbf{W}$.

The networks $\mathcal{N}_1$ and $\mathcal{N}_2$ are analyzed using standard methods for one-mode networks.

Another very important application of network multiplication is producing different networks from data tables. A *data table* $\mathcal{T}$ is a set of *records* $\mathcal{T} = \{T_k\colon k \in \mathcal{K}\}$,



Social Network Analysis, Large-Scale, Figure 21
The main two islands in ProjInst

where $\mathcal{K}$ is the set of *keys*. A record has the form $T_k = (k, q_1(k), q_2(k), \ldots, q_r(k))$ where $q_i(k)$ is the value of the *property* (attribute) $\mathbf{q}_i$ for the key k.

Suppose that the property $\mathbf{q}$ has the range $2^{\mathcal{Q}}$. For example: Authors[WasFau] = {S. Wasserman, K. Faust}, PubYear [WasFau] = {1994}, … If $\mathcal{Q}$ is finite (it can always be transformed in such set by partitioning the set $\mathcal{Q}$ and recoding the values) we can assign to the property $\mathbf{q}$ a two-mode network $\mathcal{K} \times \mathbf{q} = (\mathcal{K}, \mathcal{Q}, \mathcal{E}, w)$ where $(k, v) \in \mathcal{E}$ iff $v \in q(k)$, and $w(k, v) = 1$.

Also, for properties $\mathbf{q}_i$ and $\mathbf{q}_j$ we can define a two-mode network $\mathbf{q}_i \times \mathbf{q}_j = (\mathcal{Q}_i, \mathcal{Q}_j, \mathcal{E}, w)$ where $(u, v) \in \mathcal{E}$ iff $\exists k \in \mathcal{K}\colon (u \in q_i(k) \wedge v \in q_j(k))$, and $w(u, v) = \operatorname{card}\big(\{k \in \mathcal{K}\colon (u \in q_i(k) \wedge v \in q_j(k))\}\big)$.

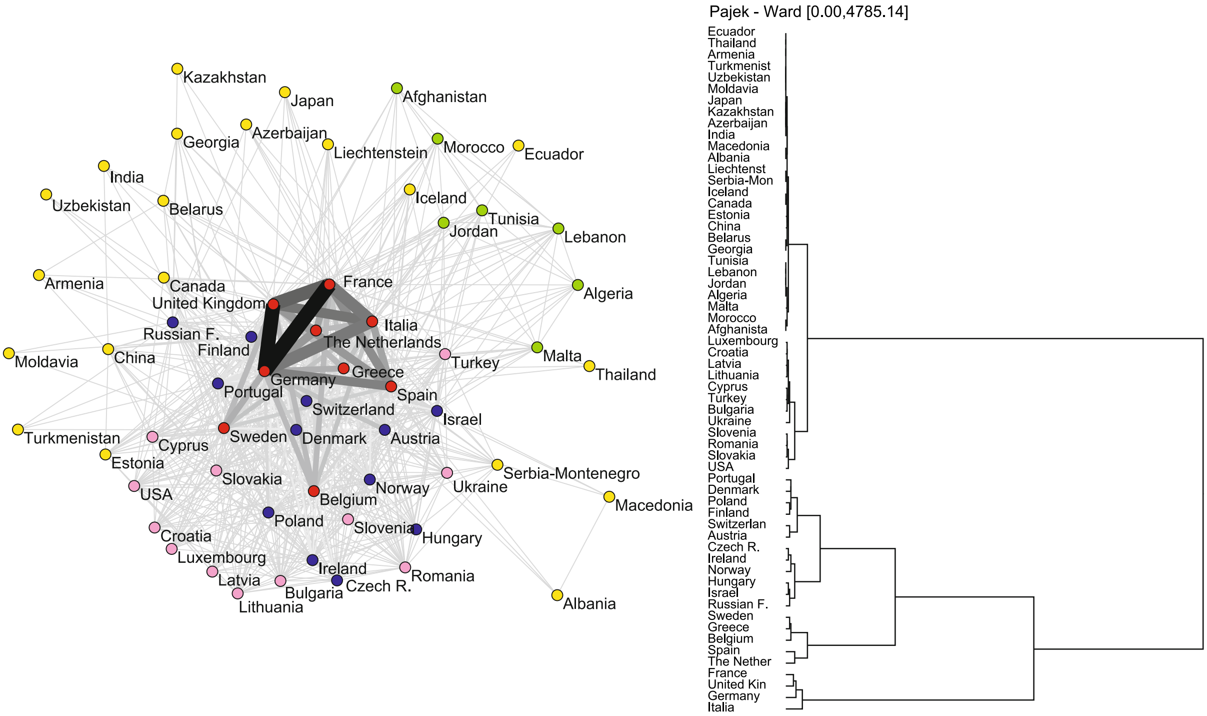
It holds $[\mathbf{q}_i \times \mathbf{q}_j]^{\mathrm{T}} = \mathbf{q}_j \times \mathbf{q}_i$ and $\mathbf{q}_i \times \mathbf{q}_j = [\mathcal{K} \times \mathbf{q}_i]^{\mathrm{T}} \star [\mathcal{K} \times \mathbf{q}_j] = [\mathbf{q}_i \times \mathcal{K}] \star [\mathcal{K} \times \mathbf{q}_j]$.

We can join a pair of properties $\mathbf{q}_i$ and $\mathbf{q}_j$ also with respect to the third property $\mathbf{q}_s$: we get a two-mode network $[\mathbf{q}_i \times \mathbf{q}_j]/\mathbf{q}_s = [\mathbf{q}_i \times \mathbf{q}_s] \star [\mathbf{q}_s \times \mathbf{q}_j]$.

For the meeting ***The Age of Simulation*** at Ars Electronica in Linz, January 2006, a dataset of EU projects on simulation was collected by FAS research, Vienna and stored in the form of Excel table. The rows are the descriptions of projects participants (idents) and columns correspond to different their properties. From this table three two-mode networks were produced: Project – $\mathbf{P}$ = [idents × projects]; Country – $\mathbf{C}$ = [idents × countries]; and Institution – $\mathbf{U}$ = [idents × institutions]; where $|\mathit{idents}| = 8869$, $|\mathit{projects}| = 933$, $|\mathit{institutions}| = 3438$, and $|\mathit{countries}| = 60$.

Since all three networks have the common set (idents) we can derive from them using network multiplication several interesting networks, such as: ProjInst – $\mathbf{W}$ = [projects × institutions] = $\mathbf{P}^{\mathrm{T}} \star \mathbf{U}$; Countries – $\mathbf{S}$ = [countries × countries] = $\mathbf{C}^{\mathrm{T}} \star \mathbf{C}$; and Institutions – $\mathbf{Q}$ = [institutions × institutions]/projects = $\mathbf{W}^{\mathrm{T}} \star \mathbf{W}$.

For identifying important parts of ProjInst network the 4-rings weights were computed and in the obtained



Social Network Analysis, Large-Scale, Figure 22
Collaboration among countries – graph and hierarchical clustering

network the line islands were determined. 101 islands were obtained, 18 of the size at least 5 (see Fig. 21). The two most important islands are: aviation companies and car companies.

In Fig. 22 the collaboration among countries is presented. For dense (sub)-networks we get better visualization by using matrix display. To determine the ordering of vertices we used Ward's clustering procedure with corrected Euclidean distance as dissimilarity measure [21]. The permutation determined by hierarchy can often be improved by changing the positions of clusters in the clustering tree. We get a typical center-periphery structure (see Fig. 23).

Note that in matrix display some details become apparent, such as the collaboration inside the peripherical group Afghanistan, Morocco, Malta, Tunisia, Lebanon, Jordan and Algeria; or a collaboration of Russian Federation with ex-Soviet republics Turkmenistan, Uzbekistan, Moldavia, Kazakhstan, Azerbaijan and Japan.

Statistical Approach

There are many properties *computed* from the network data that give us different information about it. For example:

Global properties Number of vertices, lines (edges/arcs), components; diameter; centralization; maximum core number, …

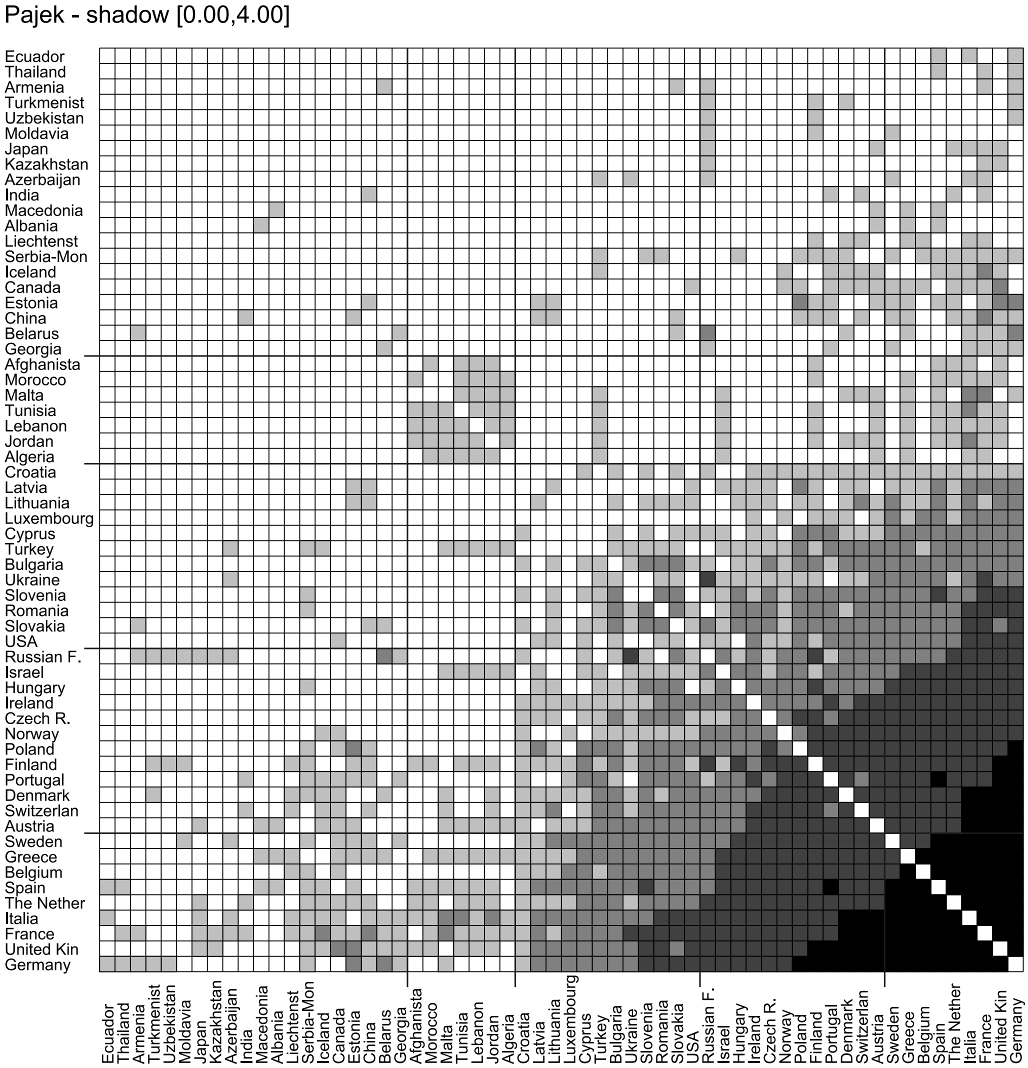
Local properties Degrees, core numbers, indices (betweeness, hubs, authorities, …). Usually we look at their *distributions* or *inspect* the values of interesting elements.

Another interesting task is searching for associations between computed (structural) data and input (measured) data.

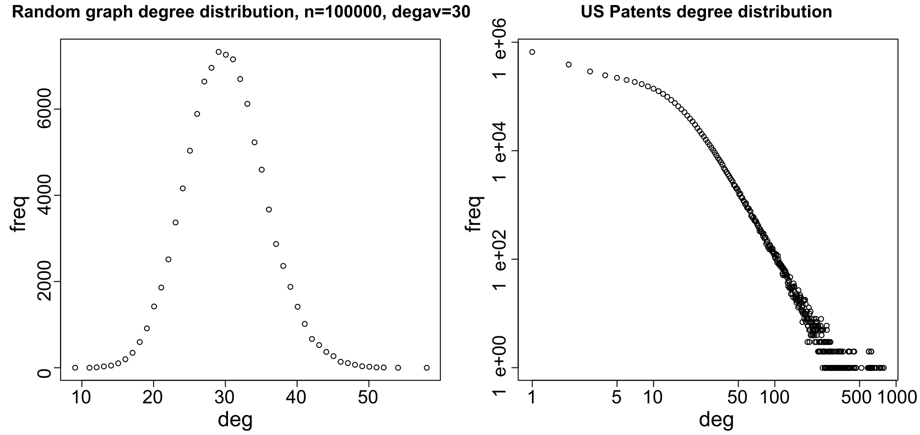
Paul Erdős and Alfréd Rényi introduced in 1959 the notion of random graph in which each pair of vertices is linked with a given probability p. The theory of ER random graphs is well developed (see [15]). Some characteristic results:

- The degree distribution is binomial (in the limit Poisson's) and most of the vertices have degree (very) close to the average degree;
- For $p \geq \frac{1}{n}$ cycles appear in the graph, and soon also the *giant component*;
- For $p \geq \frac{\log_2 n}{n}$ almost all graphs are connected.

Real-life networks are usually not random in the Erdős–Rényi sense. The analysis of their distributions gave new views about their structure.



Social Network Analysis, Large-Scale, Figure 23
Matrix display of collaboration among countries



Social Network Analysis, Large-Scale, Figure 24
Distributions: ER-random and US patents

On the left side of Fig. 24 a degree distribution in ER graph on $n = 100{,}000$ vertices with average degree $\overline{\deg} = 30$ is presented. On the right side a degree distribution for US patents citation network is presented (in log–log scale). Evidently this distribution is very far away from Poisson distribution.

In 1967 a psychologist Stanley Milgram made his experiment with letters. The letter should reach a target person. The persons involved in experiment were asked to send the letter with these instructions to his or her acquaintance that is supposed to be closer (in the acquaintances network) to the target person. The letter was sent from Boston to Omaha. The average length of the successful paths was six – *six degrees of separation*. The average path length on the internet is 19 clicks.

The networks in which the average shortest path length is small are called *small worlds*. Duncan Watts and Steven Strogatz developed in late 90-ties a procedure for construction of (random) small worlds by *rewiring* – an edge is randomly selected and one of its endpoints is attached to same other vertex. After each rewiring step the average length of geodesics is usually decreased because the rewiring creates shortcuts.

Albert-László Barabási from University of Notre Dame in 1998 analyzed several networks and noticed:

- The degree distribution follows the *power law* – the probability p_d that a vertex has a degree d equals to $p_d = cd^{-\gamma}$. In a log–log scale diagram it is represented by a line.
- In a network there exist some vertices with *large degree* (very improbable in ER graphs). These vertices link the network into a single component.

It turned out that most of real life networks (persons – e-mail, phone calls, sexual contacts (drug users, AIDS), collaboration; movie actors – playing in the same movie; proteins – interactions; words – semantic relations; ...) have such characteristics. Because for these networks their degree distribution has no natural scale they were named *scale-free* networks. For a discussion about the notion of scale-free network see [33].

The following was the first explanation (Barabási) of scale-free nature of many real-life networks:

- These networks are growing.
- In this process new vertices are added and linked with new edges to already existing vertices. The random selection of vertex to which a new vertex is attached is not uniform but follows the *preferential attachment* rule – the selection probability is proportional to the degree of a vertex.

Based on this model it can be shown the following:

- The degree distribution is the power law.
- The average length of geodesics is $O(\log n)$.
- These networks are resilient against random vertex or edge removals (random attacks), but quickly become disconnected when large degree nodes (Achilles' heel) are removed (targeted attacks).

Mark Granovetter noticed in 1973 that in social networks groups appear linked with *strong ties* [26]. They link in larger networks with *weak ties*. Also in other real-life networks vertices often form groups – the clustering coefficient is larger than in ER networks.

Several improvements and alternative models were proposed that also produce scale-free networks with some additional properties characteristic for real-life networks: copying [31], combining random and preferential attachment [37], R-mat [18], forest fire [32], aging, fitness, nonlinear preferences, ...

There are several applications of the scale-free networks theory. For example searching (Adamic et al. [2]) and spreading of epidemies (Barthélemy, Barrat, Pastor-Sattoras, Vespignani, Complex Networks Collaboratory).

For general overviews see [4,22,35,36].

Future Directions

In 2005 the support for *multi-relational* networks was introduced in Pajek. Combined with *temporal* networks it enables analysis of new kinds of networks – such as KEDS networks (*Kansas Event Data System* or *Tabari*). These networks are usually small in terms of vertices but can be (very) large in terms of lines – different interaction events among actors.

The last developed approach for analysis of large networks is adaptation of hierarchical clustering with relational constraints based on Ferligoj and Batagelj [23] to large networks. The basic idea to get a fast algorithm is to compute the dissimilarities between units (vertices) only for the linked pairs of units [11]. This approach is one of the possible approaches to analysis of spatial networks.

There are still several fields of social network analysis for which efficient approaches to deal with large networks have to be developed such as blockmodeling, probabilistic models, ...

In the near future new versions of network analysis software will appear using very large computer memories enabled by the new 64-bit computer architecture. A special challenge is development of methods and software for analysis of huge networks.

Bibliography

Primary Literature

1. Abello J, Pardalos PM, Resende MG (2002) Handbook of Massive Data Sets. Springer, Heidelberg
2. Adamic LA, Lukose RM, Huberman BA (2002) Local Search in Unstructured Networks. In: Bornholdt S, Schuster HG (eds) Handbook of Graphs and Networks: From the Genome to the Internet. Wiley-VCH, Berlin
3. Ahmed A, Batagelj V, Fu X, Hong SH, Merrick D, Mrvar A (2007) Visualisation and Analysis of the Internet Movie Database. Proceedings of the Asia-Pacific Symposium on Visualisation (APVIS2007), Sydney, Australia, 5–7 Feb. IEEE, New York, pp 17–24
4. Albert R, Barabási AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97
5. Alvarez-Hamelin JI, Dall'Asta L, Barrat A, Vespignani A (2005) *k*-core decomposition: a tool for the visualization of large scale networks. cs.NI/0504107 published in Advances in Neural Information Processing Systems 18, Canada, 2006
6. Batagelj V (1989) Similarity measures between structured objects. In: Graovac A (ed) Proceedings of International Course and Conference on the Interfaces between Mathematics, Chemistry and Computer Science, Dubrovnik, 20–25 June 1988. Studies in Physical and Theoretical Chemistry, vol 63. Elsevier/North-Holland, Amsterdam, pp 25–40
7. Batagelj V, Brandes U (2005) Efficient Generation of Large Random Networks. Phys Rev E 71:036113
8. Batagelj V, Ferligoj A (2000) Clustering relational data. In: Gaul W, Opitz O, Schader M (eds) Data Analysis. Springer, Berlin, pp 3–15
9. Batagelj V, Mrvar A (2000) Some Analyses of Erdős Collaboration Graph. Soc Netw 22:173–186
10. Batagelj V, Mrvar A (2001) A Subquadratic Triad Census Algorithm for Large Sparse Networks with Small Maximum Degree. Soc Netw 23:237–43
11. Batagelj V, Mrvar A (2007) Hierarchical clustering with relational constraints of large data sets. 6th Slovenian International Conference on Graph Theory, Bled, 24–30 June
12. Batagelj V, Mrvar A (2008) Analysis of kinship relations with Pajek. Soc Sci Comput Rev 26(2):224–246
13. Batagelj V, Zaveršnik M (2002) Generalized Cores. arxiv cs.DS/0202039
14. Batagelj V, Zaveršnik M (2007) Short cycle connectivity. Discret Math 307(3–5):310–318
15. Bollobás B (2001) Random Graphs. Cambridge University Press, Cambridge
16. Brandes U (2001) A Faster Algorithm for Betweenness Centrality. J Math Soc 25(2):163–177
17. Breiger RL (2004) The analysis of social networks. In: Hardy M, Bryman A (eds) Handbook of data analysis. Sage, London, pp 505–526
18. Chakrabarti D, Zhan Y, Faloutsos C (2004) R-MAT: A Recursive Model for Graph Mining. In: SIAM Data Mining 2004, Orlando, Florida, SIAM
19. Cormen TH, Leiserson CE, Rivest RL, Stein C (2001) Introduction to algorithms. MIT Press, Cambridge
20. Doreian P, Batagelj V, Ferligoj A (2000) Symmetric-Acyclic Decompositions of Networks. J Classif 17(1):3–28
21. Doreian P, Batagelj V, Ferligoj A (2005) Generalized Blockmodeling. Cambridge University Press, Cambridge
22. Dorogovtsev SN, Mendes JFF (2003) Evolution of networks: from biological nets to the internet and www. Oxford University Press, Oxford
23. Ferligoj A, Batagelj V (1983) Some types of clustering with relational constraints. Psychometrika 48(4):541–552
24. Freeman LC (1979) Centrality in Social Networks: A Conceptual Clarification. Soc Netw 1:211–213
25. Garfield E, Sher IH, Torpie RJ (1964) The Use of Citation Data in Writing the History of Science. The Institute for Scientific Information, Philadelphia
26. Granovetter M (1973) The Strength of Weak Ties. Am J Sociol 78:1360–80
27. Harary F, Norman RZ, Cartwright D (1965) Structural Models: An Introduction to the Theory of Directed Graphs. Wiley, New York
28. Huisman M, van Duijn MAJ (2005) Software for social network analysis. In: Carrington PJ, Scott J, Wasserman S (eds) Models and methods in social network analysis. Cambridge University Press, Cambridge, pp 270–316
29. Hummon NP, Doreian P (1990) Computational Methods for Social Network Analysis. Soc Netw 12:273–288
30. Kleinberg J (1998) Authoritative sources in a hyperlinked environment. Proc 9th ACM-SIAM Symposium on Discrete Algorithms
31. Kleinberg J, Kumar R, Raghavan P, Rajagopalan S, Tomkins A (1999) The Web as a graph: measurements, models and methods. Proc of the 5th International Computing and combinatorics Conference
32. Leskovec J, Kleinberg J, Faloutsos C (2006) Laws of Graph Evolution: Densification and Shrinking Diameters. ACM Transactions on Knowledge Discovery from Data (TKDD) vol 1, issue 1, article 2
33. Li L, Alderson D, Tanaka R, Doyle JC, Willinger W (2007) Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications. cond-mat/0501169, Internet Math 2(4):431–523
34. Mane KK, Börner K (2004) Mapping topics and topic bursts in PNAS. Proc Natl Acad Sci USA 101:5287–5290
35. Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45:167–256
36. Newman MEJ, Barabási AL, Watts D (2006) The Structure and Dynamics of Networks. Princeton Studies in Complexity. Princeton University Press, Princeton
37. Pennock DM, Flake GW, Lawrence S, Glover EJ, Giles CL (2002) Winners don't take all: Characterizing the competition for links on the web. Proc Natl Acad Sci USA 99(8):5207–5211
38. Seidman SB (1983) Network Structure And Minimum Degree. Soc Netw 5:269–287
39. Schank T, Wagner D (2005) Finding, counting and listing all triangles in large graphs, an experimental study. In: Workshop on Experimental and Efficient Algorithms (WEA). Lecture Notes in Computer Science, vol 3503, Springer, pp 606–609
40. Snyder D, Kick E (1979) The World System and World Trade: An Empirical Exploration of Conceptual Conflicts. Sociol Q 20(1):23–36
41. Snijders TAB (2005) Models for Longitudinal Network Data. In: Carrington P, Scott J, Wasserman S (eds) Models and methods in social network analysis. Cambridge University Press, New York

42. Stuckenschmidt H, Klein M (2004) Structure-Based Partitioning of Large Concept Hierarchies. Proc of the 3rd International Semantic Web Conference ISWC 2004, Hiroshima, Japan
43. Wasserman S, Faust K (1994) Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge
44. White DR, Batagelj V, Mrvar A (1999) Analyzing Large Kinship and Marriage Networks with Pgraph and Pajek. Soc Sci Comput Rev 17:245–274
45. Zaveršnik M, Batagelj V (2004) Islands. Slides from Sunbelt XXIV, Portorož, Slovenia, 12–16 May

Books and Reviews

Ahuja RK, Magnanti TL, Orlin JB (1993) Network Flows: Theory, Algorithms, and Applications. Prentice Hall, Englewood Cliffs

Batagelj V, Mrvar A (2003) Pajek – Analysis and Visualization of Large Networks. In: Jünger M, Mutzel P (eds) Graph Drawing Software. Springer, Berlin, pp 77–103

Brandes U, Erlebach T (2005) Network Analysis: Methodological Foundations. Lecture Notes in Computer Science. Springer, Berlin

Carrington PJ, Scott J, Wasserman S (2005) Models and Methods in Social Network Analysis. Cambridge University Press, Cambridge

Degenne A, Forsé M (1999) Introducing Social Networks. SAGE Publications, London

de Nooy W, Mrvar A, Batagelj V (2005) Exploratory Social Network Analysis with Pajek. Cambridge University Press, Cambridge

Knuth DE (1993) The Stanford GraphBase: A Platform for Combinatorial Computing. Addison-Wesley, Reading

Scott JP (2000) Social Network Analysis: A Handbook. SAGE Publications, London

Web Resources

Center for Complex Network Research, Notre Dame: http://www.nd.edu/networks/

Center for Spatially Integrated Social Science: http://www.csiss.org/

Complex Networks Collaboratory: http://cxnets.googlepages.com/

Internet Movie Database http://www.imdb.com/

Matthieu Latapy. Triangle computation web page. http://www-rp.lip6.fr/latapy/Triangles/

Nber: http://www.nber.org/patents/

Netminer: http://www.netminer.com/

Pajek: http://pajek.imfm.si data sets: http://vlado.fmf.uni-lj.si/pub/networks/data/

The Edinburgh Associative Thesaurus: http://www.eat.rl.ac.uk/

The Kansas Event Data System: http://web.ku.edu/keds/

UCINET: http://www.analytictech.com/

Social Network Analysis, Overview of

John Scott
School of Law and Social Science,
University of Plymouth, Plymouth, UK

Article Outline

Glossary

Algebraic models Approaches to network analysis that use algebraic methods for studying sets of points to produce positional analyses such as blockmodels.

Agent-based computational models Models of dynamic systems that focus on agent-level properties, seeing system level changes as consequences of the interaction of rule-following agents.

Balance theories A number of related theories concerning both the psychological state of consonance or dissonance found in a person's ideas and affects and the equilibrium or disequilibrium of one person's relations with another. The exploration of the mathematical principles of these states.

Blockmodeling An approach to positional analysis that uses algebraic methods to construct image graphs in which blocks represent connections or the absence of a connection between sets of points.

Graph theory A basic method for the analysis of networks in which the relational properties of the members of a set are seen in terms of points and lines. Pairwise connections among points are used to generate and explore system-level phenomena such as density and centralization.

Diffusion processes The processes through which innovations and other changes spread through a network, the flow and pace of change being determined by the structure of the network.

Scaling methods Geometrical techniques for displaying and analyzing a network as a mapping of points located in a multidimensional space.

Small-world models Network models based on graphs of relatively low density but high reachability. Originated in psychological experiments on the communicative effectiveness of interpersonal acquaintance.

Definition of the Subject

The formal idea of using a network as a tool of analysis originated in electrical engineering as a way of envisaging and modeling the flows of electricity through national power grids and in domestic and industrial settings. It was merely a short step to the application of the idea in civil engineering, where water supply, drainage, roads and railways could all be regarded as networks. Much later this idea was central to electronic and computer systems engineering. Parallel methodological innovations had occurred in areas of physical geography, where river drainage systems had been envisaged as networks of flows that could be modeled in mathematical terms, and this led to interest among human geographers about the potential for applying the idea. The idea that social relationships could be seen as forming a network or reticule of relations that are capable of analysis in formal, mathematical terms suggested novel ways of considering social phenomena and promised insights that were unavailable through alternative approaches.

Social network analysis first emerged as an approach distinctive from statistical analysis within anthropology, social psychology, and sociology. It went on to influence theoretical and methodological developments in many other areas of social sciences. In all these areas, researchers took up mathematical models from graph theory, which they used to investigate structures of relations. Researchers who were working in the area of social network analysis soon began to broaden their framework of analysis by exploring the implications of other mathematical models, such as algebraic set theory, scaling methods, and statistical and stochastic methods. In this way, social network analysis became linked to the larger intellectual debates going on in other areas of the natural and social sciences where these same methods were being drawn upon, and a number of social network analysts have contributed to the development of these and other mathematical models. Social network analysts have promoted and undertaken the application of these ideas in many areas of sociology, most notably in economic sociology, political sociology, the study of family and kinship, community studies, the investigation of world systems, and numerous other areas.

Investigation into social networks has now become far more fashionable than ever before. However, it is in the nature of intellectual fashions that much of the new interest comes from those who are least familiar with the area. This is certainly true of social network analysis, where many newcomers were unaware of its long history. They were influenced, in particular, by the exaggerated claims to intellectual novelty that were being made by some physicists who have advocated random-network and small-world models as the basis of what they call a "new science of networks". Indeed, many people have assumed that social network analysis began de novo with these models and they know little about the applications already made using other models. Despite their exaggerated claims, the work of the physicists has provided us with some powerful new models, and these have had a major impact beyond the specifically social sphere – this fact will be apparent from considering chapters in a number of the sections in this *Encyclopedia*. However, the physicists have not provided a complete replacement for social network analysis, as they sometimes imply, but have offered a broadening of its range and concerns.

This overview article will discuss the development of social network analysis and will sketch the various mathematical models that have been used to explore the nature of complex social systems. It begins with an overview of the history of social network analysis and its main applications. This is followed by a consideration of graph theory, the earliest and principal approach to the study of social networks. The following sections look at later and more advanced methods: algebraic methods and blockmodeling, scaling methods and visualization techniques, statistical methods, agent-based computational methods, and finally the small-world methods of the physicists. The latter discussion will place the new social physics in the context of earlier discussions and it will show the variety of mathematical models that have a part to play in contemporary social network analysis. The various methods and models are considered in outline and without technical detail. The various specialist chapters included in this and other sections of the Encyclopedia cover these methods in greater detail.

The Development of Social Network Analysis

Social network analysis has a long history, though it appeared under its specific name only far more recently (see Chap. 2, p. 23 in [51]). Although structural thinking was central to classical sociology, the intellectual roots of a specifically network perspective can be traced back to those sociologists of the late nineteenth and early twentieth centuries who used the idea of a network of relations as a metaphor for understanding social relations as *structures* of relations. Georg Simmel and other German sociologists, in particular, wrote extensively on the "interweaving" of actions to form complex configurations or "webs" of connection. Their metaphorical usage of the idea drew explicitly from the textile industry and its ideas of the weft and warp of woven fabrics. The first proper usage of the actual word "network" in a sociological context, however, was in the work of the anthropologist Alfred Radcliffe-Brown [45], who used it to suggest a more formal focus for structural analysis. Simultaneously with this, a number of social psychologists working in the area of child study and field theory began to construct diagrams of social relations that they gradually formalized as "sociograms" to show the chains of positive and negative connections that they were finding among the individuals in their experimental studies.

The crucial social psychological work on social networks was that of the psychotherapist Jacob Moreno [41]. Following his arrival in the United States he began to combine psychodrama with studies of the influence of friendship choices on personality development among schoolchildren. His key innovation was to make use of sociograms to depict the friendship patterns that he was able to discover within small groups such as a school class or a neighborhood peer group. Moreno made many intellectual contacts, most importantly with the *Gestalt* psychologist Kurt Lewin, who had himself recently migrated to the United States. Moreno and Lewin exchanged many ideas – and the nature of the exchange became a matter of contention between them in later years. Lewin soon began to move towards a more social form of psychology that stressed the importance of looking at organized patterns of group relations in particular spheres of activity as forming systemic wholes. Such wholes could be seen, by analogy with electromagnetic and gravitational fields, as social fields through which influences are able to flow from one area or part to another. The aim of Lewin's field theory was to map and to explore the "dynamics" of these fields of interpersonal influence.

Both Moreno and Lewin used diagrammatic representations of social relations, but it was Moreno who invented the classic form of the sociogram. A sociogram consists of points, representing individuals, and lines, representing the social relationships among them. Moreno argued that these patterns of points and lines – later described more formally as "vertices" and "edges" – could be described as "social configurations" in which could be charted such things as the direction of friendship choices and their intensity. This work came to be referred to as "sociometry", and a journal with that name was founded in 1937. Through the work published in this journal, the ideas of Moreno and Lewin on configurations of social relations in small groups gradually gave rise to a more mathematically oriented approach that went under the name of "group dynamics". It was in this school of thought that ideas from graph theory – first formulated by René König [33] – began to be employed from the late 1940s, these ideas becoming the basis of a more formal and rigorous study of the properties of the social networks of small groups. By treating the sociogram as a graph, it proved possible to chart dyadic, triadic, and more complex structures and to analyze the states of "balance" and imbalance that existed among positive and negative relations within each network [3,13,28].

It was during the 1930s and 1940s that anthropological and sociological work also began to move in a similar direction. In 1929, the anthropologist W. Lloyd Warner, a former colleague of Radcliffe-Brown, had moved to Harvard University in order to work with the psychologist Elton Mayo on a series of experimental studies of work behavior. Their aim was to use the techniques already employed by anthropologists in studies of small tribes and villages, but to carry out new studies of small groups in the factory and office settings of urban societies. Influenced also by the early system theory constructed by the physiologists Lawrence Henderson and Walter Cannon, they began a series of empirical studies at the Hawthorne electrical works in Chicago. The so-called Hawthorne studies became a milestone in social research – not least because of their significance in developing the idea of the social network.

The Hawthorne studies [47] had begun as management-run investigations into worker efficiency and its relationship to the physical conditions of work. The early experiments examined such things as the consequences of changes in levels of heating and lighting and the provision of work breaks. The studies turned from physical to social factors under Warner and Mayo, and Mayo famously reported that the crucial factor in increasing worker productivity had been their involvement in the experiment itself. The paradoxical results showing increased productivity even when physical conditions worsened were interpreted as showing the high levels of satisfaction that workers felt simply from being part of the management experiment: it was the first time that anyone from management had shown any direct interest in them. Adopting more anthropological methods of observation in the bank wiring room, the researchers documented the gradual building of informal work group relations and forms of social solidarity among those in the experimental group. These informal relations ran in parallel with the formal relations of the managerial organization chart, and the researchers used sociograms to chart the informal social networks among workers.

The sociograms drawn by the Hawthorne researchers allowed them to identify "cliques" and other sub-groups within the larger group, and Warner went on to investigate this kind of group structure further in a series of investigations into community relations in the New England city of Newburyport (referred to as “Yankee City” in [59,60,61,62]), the Mississippi city of Natchez (“Old City” in [15]), and the Indiana city of Morris (“Jonesville” in [58]). In this work, Warner and his colleagues used network ideas to explore the formation of cliques, “crowds”, and “circles” among those they studied, and they examined these in relation to the more formal structures of economic and political life. Their investigations identified these social circles through a rough and ready use of the Venn diagrams of set theory and the methods of matrix display. Using these simple methods, they were able to identify such things as class divisions in clique membership and the effects of ethnicity on the structure of network relations.

It was during the late 1940s and early 1950s that a significant consolidation of the social network approaches that had been developing began to be forged. It was to be another decade, however, before a recognizable tradition of social network analysis was properly established. George Homans, a colleague of Warner and Mayo at Harvard, attempted a synthesis of the anthropological and sociological ideas as the basis of an alternative sociology to the system theory of Talcott Parsons [30]. Homans re-analyzed some of the earlier data using rudimentary techniques of matrix rearrangement, his aim being to bring out the clique structure more clearly. Homans’ method of manually “reshuffling” the 0s and 1s in a matrix to bring out blocks of high and low density was a pioneering move towards the more formal methods of blockmodeling that developed during the 1970s. Homans himself interpreted his data on social relations in terms of the frequency and duration of the relationships involved and went on to develop rational choice models of action to explain the findings [31].

The early 1950s also saw the development of more systematic anthropological work on social networks in Britain. The Manchester University anthropologist Max Gluckman established the Rhodes-Livingstone Institute in Northern Rhodesia (now Zambia) to study African social structures in village and town settings. An early associate of the Institute was John Barnes, who undertook independent fieldwork in Bremnes, Norway. It was here that he recognized the importance of the network metaphor – famously while observing the repaired fishing nets of the local fishermen – and he advocated the study of whole communities as networks of points and lines [2]. Elizabeth Bott, a student of Lloyd Warner and then working at the Tavistock Institute (associated with the Michigan Research Center for Group Dynamics), took up this metaphor in her own investigations into the kinship networks of London families [6,7]. Because of her position at the Tavistock, she was encouraged to link this to the work of Moreno and to the early field theory. This work was systematized by Siegfried Nadel and Clyde Mitchell. Nadel [42] aimed to establish a mathematical basis for social network research, but his early death prevented him from carrying this further. Mitchell [39], however, made the network idea the basis for a series of studies at the Rhodes-Livingstone Institute and set out a number of conceptual innovations – density, multiplexity, durability, direction, reachability and so forth – that echoed and took further the ideas of Homans.

By 1969, then, a clear body of network ideas had been established. in anthropology and social psychology, and this was coming to be organized around the mathematical theory of graphs. It was at this point, however, that an alternative framework began to develop in the United States, and this eventually subsumed the earlier approach in a comprehensive framework of social network analysis. John Boyd, François Lorrain, and Harrison White had begun to explore some aspects of Lévi-Strauss’s analysis of kinship structures using algebraic set theory [72], while Edward Laumann [34] and Joel Levine [35] began to use multidimensional scaling to investigate social structure. White moved to Harvard University and brought together a strikingly original group of young researchers who were interested in using mathematical methods to explore aspects of social networks and social relations.

One of the earliest products of this group was Mark Granovetter’s [26,27] influential argument about “the strength of weak ties” – the first, and most important, of the counter-intuitive results to come out of this more sophisticated form of network analysis. Granovetter’s argument held that the chances that a person will receive useful information from those in his or her social network is greater when a person has acquaintances rather than close and intimate friends. One’s intimates tend to have access to the same pool of information through their dense pattern of mutual connections. Acquaintances, on the other hand are connected to quite diverse and distant social networks and so can communicate a much greater range of information. Paradoxically, then, it is the weak acquaintanceship ties that are the “strongest” or most important carriers of information. Having a small number of acquaintances can be more useful in a job search than having a large number of intimates.

The ability of White’s group to pursue this program of social network analysis was made possible by the advances that were being made in computing technology and that were then becoming available in software programs. Earlier work had used manual techniques and so had been limited to relatively small groups. Mainframe computers and Fortran programming made possible the analysis of much larger data sets and the use of more powerful mathematics. As the members of the Harvard group spread to Graduate centers across North America, their work became a major influence within sociology and encouraged the global spread of formal social network analysis. Barry Wellman and Steven Berkowitz were particularly active in setting up the *International Network for Social Network Analysis* (http://www.insna.org/) and this became the basis for the founding of the journal *Social Networks* and for their own powerful syntheses in social network analysis [4,71] (see also [32]). Powerful methodological ideas and sociological applications were developed in the United States by Ronald Breiger, Ronald Burt [10,11], Patrick Doreian, and others. Social network analysis was pursued in Canada by Barry Wellman and Peter Carrington, by John Scott in Britain [51], by Rob Mokken and Frans Stokman in the Netherlands (the developers of the *Gradap* program used in [29]), and by Charles Crothers and Malcolm Alexander in Australia. Stanley Wasserman led the way in producing a standard textbook for social network methods [64] and other consolidating texts appeared [24,37]. By the 1990s, social network analysis had become one of the most strongly established research specialisms in Sociology, with a strong intellectual base, advanced methods and texts [12,18], and a comprehensive range of exemplary studies.

During the 1990s, however, a rather paradoxical development occurred. The exhaustion of many of the conventional research programmes in physics had led a number of theoretical physicists to search out new areas in which to apply their theories. In 1998, Duncan Watts and Steven Strogatz published a paper [68] that revisited some of the ideas on random networks that had grown out of Stanley Milgram's work on "small worlds" [38,56]. This caused a great stir among physicists and led Albértó-Laszlo Barabási to propose the building of a "new science of networks" with the potential for applications across the social sciences [1]. In apparent ignorance of the prior existence of social network analysis, he claimed that it was only during the 1990s – and thanks to physicists' models – that investigators had become aware of the fact that social networks have structures and show orderly patterns of development.

This claim startled and dumbfounded most sociologists working on social networks, for whom the existence of social structure was a fundamental axiom. They resented the ignoring of what they had already achieved and the misrepresentation of sociological understandings of the social world that it involved. Despite its shaky historiography, the arguments of Bárabasi and other social physicists had a major influence outside the discipline and among those in sociological specialisms that were, so far, untouched by social network analysis. Works by Duncan Watts were particularly important [66,67]. The new social physics of networks was especially well-received in the popular and scientific press and resulted in a number of popularizing works (see, for example [8]).

It is, nevertheless, true that the work of the physicists has certainly been responsible for a substantial growth of interest in social network analysis during the last decade. The reason for the appeal of this new work lies in its promise to provide a dynamic model of social networks, rather than the more static ones of the past. While the claim that social network analysis lacks a concern for dynamic processes is, perhaps, overstated, it is certainly the case that the new methods do seem to offer the possibility of a nuanced account of structural change in social networks. While much of the work is still presented without any recognition of prior sociological contributions, sociologists, for their part, are beginning to discuss how these ideas might enrich social network analysis. The arguments about network dynamics are, in fact, converging with some of the work on agent-based computational models already being undertaken within Sociology.

A very important issue that has arisen in the use of the mathematics of graphs is that of whether it provides a substantive *theory* or simply a *method* of analysis. Some researchers hold that the mathematical theorems that can be derived from graph theory can be directly translated into substantive theorems concerning particular domains of application. According to this point of view, sociological laws can be derived directly from mathematical laws. Those who see social network analysis as a method – perhaps a majority of those working in the area – hold that mathematical theorems provide simply the formal constraints to which social processes must conform: a knowledge of mathematics does not, in itself, obviate the need to study social phenomena empirically in order to understand the implications of these theorems for particular types of social relations.

The remainder of this overview article will set out the main intellectual strands within contemporary social network analysis. These areas are discussed in greater detail in the specialist articles in this and other sections of the *Encyclopedia*, and their mathematical bases will not be presented in any detail. The overview will concentrate on the general principles involved in each area and on the ways in which they relate to each other.

Graph Theory and Ideas of Balance

Graph theory, as already noted, is that branch of mathematics that had the earliest influence on the development of social network analysis (see ► Social Network Analysis, Graph Theoretical Approaches to). In graph theory, a graph or network is simply a collection of points (vertices) connected by lines (edges). Graph theory comprises a set of mathematical axioms and derivations that describe the actual and potential patterns formed by these points and lines. This basic model of a graph can be applied widely in many substantive domains: electrical wiring networks, river drainage networks, road and transportation networks, and computer networks have all been investigated using graph theory.

A basic graph consists of simple points and lines, but the lines in a graph can be assigned positive or negative attributes, directions, or numerical values, and the attributes of points can be recorded. Complex mathematical theorems can be constructed for each of the different types of graph. Some of the earliest work to use these ideas focused on the question of "balance" in graphs. Balance is a term that describes the pattern of signs and numerical values in a graph in terms of the reciprocity and transitivity of the relations. A graph of friendship relations, for example, might be unbalanced if two people who each like one another have opposite relationships to a third person. If, on the other hand, all three people liked one another with a similar intensity, the graph would be balanced. Balance theorists have analyzed complex graphs in terms of such "triads" of relations. By producing a census of triads and computing the balance or imbalance in each triad, it is possible to arrive at a calculation of the overall balance in the graph as a whole. Such ideas of balance have generally been assimilated to the idea of equilibrium analyzed in classical mechanical models. A graph that is in a state of disequilibrium, therefore, will be marked by tensions and stresses that push and pull it towards a state of equilibrium. Where a social network is mapped as an unbalanced graph, it can be hypothesized that the participants in the network will experience these stresses as push and pull factors on the social relations that they maintain with each other. The overall network can, therefore, be seen as a dynamic field of causal influences.

Much of the work that has been carried out using the idea of balance has considered focal actors to be the points of reference, and much of the early work that applied graph theory pursued such "ego-centric" ideas. From this point of view, the chosen actor is seen as a point that is "adjacent" to certain other points that comprise its immediate "neighborhood". Each point can, therefore, be given an adjacency score. Actors vary in terms of the number of others with whom they are connected – the number of friends, number of enemies, number of political associates, and so on. People or groups may, therefore, be more or less connected into their neighborhood of others and so they may be compared in terms of their adjacency scores. The early work of Moreno had identified school children as sociometric "stars" or "isolates" according to the number of friends they had. The stars were the especially popular class members, while the isolates had few or no friends in the class. The adjacent others may have more or fewer connections among themselves: my friends may or may not be friendly with each other, my business associates may or may not do business with each other, and so on. Thus, a basic ego-centric measure is the "density" of an actor's neighborhood. This measure is the number of links that actually exist among immediate contacts expressed as a proportion of the number of possible contacts that they could, in principle, sustain among themselves. Where a large proportion of connections intersect, the actor's neighborhood has a high density.

Researchers recognized fairly early on the importance of going beyond such ego-centric measures to more global or socio-centric measures of the network as a whole. In considering a village, for example, a researcher may want to know not only the distribution of individuals according to the densities of their friendship relations, but also the density of friendship connections in the village considered as a whole. Such measures of density would indicate the state of social cohesion or social solidarity in the network. The basic ego-centric measure of density can easily be extended to the network as a whole – if appropriate data are available – by computing the proportion of possible links within the village that are actually established. Calculations of density can also be made for large social networks that must be studied through sample investigations. If it can be assumed that the sample is representative, then the distribution of adjacency scores for sample members can be used to compute the overall density of the network.

Density has become one of the most widely used measures in studies that employ graph theory and it has been used to indicate the changing character of social solidarity and the state of "community" that exists in a group. Wellman's work [69,70] has demonstrated that growing amounts of geographical mobility have not led to a commensurate decline in community but have resulted, instead, in the transformation of patterns of solidarity. People in large cities, he shows, are likely to interact less frequently and less intensely with numerous others on a day to day basis as they move to live and work in other parts of the city. Thanks to cars, the telephone, and the internet, however, they are able to maintain more connections, but on a less frequent and less intense basis.

Particularly important measures of global connectedness within a graph are based on the "distance" from one point to another and the corresponding "centrality" that points have in the overall pattern of connection. The distance between two points in graph theoretical terms is simply the number of lines that must be traversed in moving from one point to another. Thus, two adjacent points stand at distance 1 from each other, while two points that are connected through a common neighbor are at distance 2 from each other. It is generally assumed that influence and communication within a social network attenuates with increasing distance. Nevertheless, the rate of attenuation and the particular level of distance that it is reasonable to regard as constituting a socially significant connection are quite variable from one type of network to another. Distant friends of friends, for example, may be insignificant, while distant financial donors to a political party or candidate may be more significant. The criteria of significance, it will be apparent, rests always on a sociological judgment and never on a mathematical measure alone.

The principal use of distance measures has been to assess the relative centrality or peripherality of points in a graph. A point whose average distance to all other points in the graph is low can be regarded as occupying a central position in the network as a whole, while a point with a very high average distance from all others is peripheral to it. Applications of these ideas in studies of interlocking directorships – the lines created when company directors sit on two or more company boards – have made great use of measures of centrality. They have to documented the existence of "hubs" within large spheres of adjacent companies. Such hubs – typically banks or other financial institutions – are the focal centers for business decision-making and for the allocation of credit at the level of the intercorporate network as a whole [50].

Users of graph theory have often distinguished between measures of centrality based on overall "closeness" and an idea of centrality based on "betweenness" [22]. A betweenness measure of centrality is one that is based on a calculation of the probability that any particular point lies on a path between any two other points. This latter measure can be used to operationalize the idea of the broker or intermediary, a person who may not be central in the network as a whole, but is an important point of contact between parts of the network.

A further extension of the ideas of distance and point centrality has been the attempt to measure the overall centralization of the graph as a whole. A measure of centrality grasps the extent to which a whole graph is organized around a small number of central points that constitute its structural center.

The early work of Warner and his associates was concerned with the formation of cliques within networks. Their use of this idea of the clique was, however, relatively loose and essentially commonsensical. Developments in graph theory have since made possible the construction of a number of structural concepts for sub-graphs that are both precise and distinct from each other. The most straightforward of these sub-graph concepts is that of the "component". A component is a maximally connected sub-graph, the set of all those points that are connected, directly or indirectly, by a continuous sequence of lines. A graph will typically consist of a number of such components, together with a variable number of completely isolated points. Component boundaries may, therefore, indicate the boundaries to the flow of communication, influence, or resources within the corresponding social network. A mapping of the size distribution of components will show the extent to which these flows are fragmented. It is also possible to map the internal structures of the various components by taking account of the sign, value, and direction of the constituent lines in order to uncover any blockages or distortions in the flow of communication, influence, or resources. Using a measure of value, for example, the nesting of intensely connected components within more weakly connected components can be disclosed. The "slicing" techniques available in many software packages allow contour maps of component structure to be built, the resulting map showing the peaks of intensity and the troughs or valleys that separate them.

Component structure can also be studied through an investigation of the "cycles" that it contains. A cycle is a path that returns to its own starting point, and a cyclic component consists of a set of intersecting cycles. A component may consist of a number of such cyclic components that are connected only through non-cyclic "bridges".

Cliques, as defined in graph theory, are sub-sets of points in which every possible pair of points is connected at a specified distance. Thus, a 1-clique is a sub-set of point in which all points are directly connected to each other. A 2-clique, on the other hand, is a sub-set of points that are connected through common neighbors. There is, therefore, a whole family of clique concepts based on varying distance measures. These are generically referred to as n-cliques and any large component in a graph will typically consist of many overlapping *n*-cliques, as well-connected points will tend to be members of large numbers of cliques. An analysis of cliques must always specify the value of n that will be used. Once more, the choice of a value for identifying cliques is not a matter that can be determined on mathematical grounds alone. The investigator must always decide on the sociological meaning – if any – that can be given to particular values of n.

This basic idea of the clique has been extended into a variety of related ideas, most notably called clans, clubs, plexes, and circles, each of which differs according to the precise criterion that defines sub-set membership. These various sub-graph and sub-set concepts are particularly important in studies of diffusion processes in social networks as they define the "obstacles" and barriers that prevent the smooth diffusion of ideas or resources. This was the insight that lay behind Granovetter's [26] recognition of the strength of weak ties: it is the possession of "weak" connections that bridge a person into sub-graphs that would otherwise be unable to transfer information to the person.

Most uses of graph theory in social network analysis have employed one-mode data. That is to say, they have transformed the initial bipartite or two-mode data into a simpler form. An initial incidence matrix showing, for example, the memberships of people (as columns) in particular organizations (as rows) is transformed into a column-by-column adjacency matrix of people and a row-by-row adjacency matrix of organizations. Thereafter, the analyses of the people and the organizations proceeds separately (see ► Social Network Analysis, Two–Mode Concepts in). Recently, however, attempts have been made to develop graph theory approaches to the analysis of two-mode data, treating the bipartite data as if it were one-mode data. The adjacency matrices are, in effect, seen as sub-matrices of a larger adjacency matrix in which people and organizations, for example, comprise *both* the columns and the rows. Once the larger matrix is generated, many of the conventional graph theory measures and visualization techniques can be applied to it. Not all such measures, however, are sensibly or usefully seen in two-mode terms.

Many of the more complex graph theoretical measures for whole graphs can be difficult, if not impossible, to use in large-scale networks where there is incomplete data. This is typically the case where a sampling methodology has been followed (see ► Social Network Analysis, Estimation and Sampling in).

Some network parameters can be estimated with sample data. If information on the neighborhood or neighborhood density of individuals are collected from a sample, for example, it is straightforward to compare estimates of these measures for the whole population. It would also be possible to estimate the density of the network as a whole from such sample data, as density is an adjacency-based measure. So long as the normal questions of sample size and representativeness are considered, such measures pose few problems. Matters are more complex, however, with many global network measures, where the structure of relations is more likely to be lost in the process of sampling. It is, for example, impossible to measure the number of separate components in a network using sample data. Some ingenious methods of estimation have, however, been suggested as being useful for such measures, so long as appropriate sampling designs are used.

An alternative to sampling is the adoption of explicit methods of data reduction that allow large-scale networks to be analyzed more easily. As Vladimir Batagelj shows in ► Social Network Analysis, Large-Scale, it is especially useful to have methods of visualization for such networks. The *Pajek* program [16] was designed specifically for this purpose. Large networks are those with thousands, or perhaps millions, of points and the computing time required for such networks has been a barrier to studying them. Batagelj uses recursive decomposition methods to divide a large network into several smaller ones. This can be achieved by the clustering of points or lines, compressing them into compound points or lines. Varying forms of reduction can be achieved by adopting different criteria for reduction: for example, collapsing all members of a component onto a point, or collapsing all members of a clique into a single point. Such reduced graphs can then be analyzed with the familiar techniques of graph theory. He also introduces a number of concepts specific to reduced networks. The concept of an island, for example, refers to peaks at specified cut-off levels on a measure of the properties of the compared points or lines.

Diffusion Processes

A key issue in sociological studies of influence has been the ways in which innovations and other social changes are diffused through the social relations that people have established. Similar processes of diffusion were investigated in geography and anthropology, and a number of investigators began to explore mathematical models of diffusion, that had already been begun to be applied in epidemiological studies of the spread of disease through a population [48,49] (see also [14]). Everett Rogers documented the existence of an S-curve describing the pace of innovation: a slow initial uptake was followed by more rapid adoption, and then a tailing-off as diffusion produced a relative saturation of the population and correspondingly fewer people available to take up the changes. The key parameters were the speed of the initial take-off and the speed of subsequent adoption.

Contemporary approaches to diffusion [57] begin from measures of spatial autocorrelation to investigate the speed of the spread of innovations (see ▶ Social Networks, Diffusion Processes in). Spatial autocorrelation is the correlation of a variable with itself measured spatially. That is, its dependence on neighboring points. Such models measure the extent to which changes spread through contiguous regions of space. Within networks, this idea is generalized to analyze contiguity in social network terms. It is in this way that a "contagion" model of social influence can be operationalized. Network parameters can be weighted in various ways in a model, taking, for example, direct links as more significant than indirect links, and predictive conclusions about the temporal and spatial spread of changes can be drawn. As Tom Valente shows, event history analysis has been used to allow more sophisticated temporal measures of diffusion, and this has become a key area of innovation within the area of diffusion studies. Such work is also now developing closely with the agent-based computational models discussed below.

Algebraic Models and Blockmodeling

The early work on network structures that was carried out by Harrison White and his associates [36,72] made use of algebraic ideas to model graphs of social relations. Their argument was that algebra provided a means for analyzing the properties of sets of points. This work laid the basis for what came to be known as blockmodeling. This is a way of exploring the relations among structurally defined positions in social networks, the positions being understood as "blocks" or sets of points with certain common relational attributes that make them, in crucial respects, structurally equivalent to each other [5,73]. This form of positional analysis has subsequently been elaborated and enlarged in a variety of ways, resulting in the production of a whole family of models for network structure.

An algebraic representation of a social network involves identifying a set of points that is defined by the specific relation that unites its members [43] (see ▶ Social Networks, Algebraic Models for). Thus, a network is seen as consisting of a set of ordered pairs. This approach can easily be generalized to explore situations in which more than one relation is involved and units can be understood as connected into multiple networks, each of which is defined by a particular type of relationship. From this point of view, a multiple network is a set of sets, and each set is defined by a "primitive relation" or generator that contributes to determining the overall network structure. Therefore, the individuals or groups that comprise the points in a social network are treated as standing in compound relations to each other. Paths in the network may be composed from lines that represent a variety of different relations. The basic step in identifying the structure of such a complex network is the multiplication of the initial binary relations to convert them into binary compound relations. The structure of relationships among all possible paths in a multiple network is termed a partially ordered semigroup and this comprises the underlying relational structure of the network. As with graph theory, this algebraic method allows such global properties of social networks to be linked to the local, egocentric properties of specific units within the network. Thus, the immediate environment of any particular member of the network can be seen as a "local role algebra" that represents the role set of the focal individual or group. Such an algebra defines the vectors that slice through the whole network and so summarize an ordering of the constituent relation vectors in which the individual point is involved.

The basic elements of blockmodeling are based on the ideas first set out by the anthropologist Nadel, who held that social network analysis must take account of social positions and their associated roles rather than simply with individuals [42] (see ▶ Positional Analysis and Blockmodeling). Graph theory is especially well-suited to the identification of individual-level phenomena and the characteristics of the social groups formed by individuals. It is much less well-suited to the identification of structural positions. For example, a large number of individuals may each occupy the socially determined position of "father", but a graph theoretical analysis of each individual male adult's connections to children will not easily identify this shared social role. This is because graph theory searches for clusters of connections based on distance measures. Blockmodeling, on the other hand, starts out from a recognition that individuals that occupy the same social position are "equivalent" or "substitutable" in some way. Thus, each father relates in similar ways to particular children, but the fathers do not all relate to the *same* children. Positions and their associated roles are enduring structural regularities among sets of points and differ markedly from cliques, components and other such subgraphs.

The typical strategy for constructing a blockmodel is to apply an algorithm to a bipartite, two-mode, matrix of individual-level data in order to reduce it to an "image graph" that comprises blocks of equivalent points. Though typically employing two-mode data, blockmodels can be produced for one-mode data and, with greater difficulty, for three-mode data. Simple binary data have been most widely used in blockmodeling, with valued data being handled by slicing procedures that partition the matrix into 1s and 0s on the basis of a chosen threshold of significance. Attempts are being made, however, to extend blockmodeling directly to valued data. Approaches to blockmodeling vary from each other in terms of the particular criteria that are used in the algorithm. The earliest approach – the CONCOR method of White and his associates – used iterated correlation measures to produce the image graph. In this method, repeated correlations computed on the rows and columns of the initial matrix converge towards a rearranged matrix that shows blocks of 0s and blocks of 1s. The blocks of 0s – "zero-blocks" – are the structural holes in the network that divide the various social positions (the blocks of 1s) from each other. Later work associated with Ronald Burt [9,10], has used simpler clustering approaches that are based on Euclidean distance measures calculated from path distances among pairs of points. Burt's method also allows the construction of an image graph for the identification of blocks of equivalent positions, and Burt terms this a social topology.

Applications of these ideas to intercorporate business connections has shown that blocks in networks of shareholdings and interlocking directorships can be seen as comprising structurally similar agents within the economy. The blocks uncovered in such an analysis are the dominant investors and subordinate enterprises that stand in similar relations to each other but may have few mutual or reciprocal connections among themselves. They are social positions but do not necessarily constitute social groups. The key text on blockmodeling has been produced by Doreian and colleagues [19].

The construction of an algebraic representation of a social network comprises a depiction of the "deep structure" of the network in the sense that was intended by Lévi-Strauss and many structuralist writers. The method allows a formal construction and representation of the deep structures that have, more typically, been identified through more purely interpretative means. As in these structuralist approaches, a key aim of algebraic network analysis is the identification of isomorphic subgroups and local role algebras across different empirical networks that can be compared for commonalities of structure. This comparative approach allows the formulation of statements concerning generic processes in social systems, modeled as lattices. Hierarchical structures, for example, may be common to a number of different empirical domains and may exhibit similar formal properties, regardless of their specific empirical content. Similarly, empirically variant role sets may exhibit a similarity of underlying structure, such as patterns of conflict or cooperation with different categories of others.

This approach does not, however, reduce the variation found in actual social networks to invariant mathematical properties. There can be no derivation of a network *theory* directly from a mathematical theory. As Pattison (see p. 135 in [43]) argues, algebraic approaches have sacrificed mathematical power in order to reflect more adequately the social relationships in the particular field under investigation. Algebraic methods have tended to be used – as shown in the discussion of blockmodeling – as methods of data reduction, as means for decomposing or fragmenting networks into simpler structural components. The construction of empirically relevant theories from these components is far more difficult and has, so far, rarely been pursued.

Scaling models and Visualization

Both graph theory and algebraic approaches offer rudimentary techniques for the visualization of social networks (see ► Social Network Visualization, Methods of). Graph theory, for example, is based around the idea of the sociogram of connected points. For any but the smallest networks, however, sociograms become a confused jumble of cross-cutting lines, and a rearrangement of points aimed at reducing the amount of overlap can actually destroy any visual representation of the structure of the network by distorting the distances between and the relative locations of the points. A blockmodeling approach to large networks reduces the amount of data to be handled and allows simple image graphs of the relations among blocks to be drawn, but this is limited to positional analysis and loses sight of the details that are present in the relations among the individual points. In both forms of analysis, therefore, social network analysts have tended to rely on the purely mathematical manipulation of large networks and many have almost abandoned the attempt to visually display whole complex social networks. The desire to recapture the simple visual impact of the sociogram, however, has motivated a number of researchers to investigate ways of drawing network diagrams that retain the spatial sense inherent in relational data. Unsurprisingly, perhaps, the techniques discovered have followed the methods used in cartography to represent physical landscapes on the page of a map or atlas.

The earliest and most influential visualization techniques drew on multidimensional scaling (MDS), as this was seen as a method for mapping data that retained its fundamental metrics of space and distance. Essentially similar techniques are those of principal component analysis or factor analysis. Instead of calculating distance simply by the number of lines connecting two points – "path distance" – MDS approximates to the measures of physical distance used in conventional spatial thinking. MDS techniques convert raw data on social networks into Euclidean distances. In this conversion, points are regarded as "close" to each other by virtue of their position in the overall configuration of relations. This is typically measured on the basis of a proximity measure such as the frequency of contact between two points or the intensity of a relationship between them, and these are usually given a strictly metric form by computing correlation coefficients. For example, two points with identical patterns of connection are perfectly correlated with each other, and decreasing similarity is reflected in declining closeness. Using such similarity (or dissimilarity) data, MDS procedures search for the best fit within a Euclidean space of specified dimensions. Once a solution is obtained it can be displayed on the page and rotated into an informative orientation that discloses the meaning of the dimensions that have been identified in the analysis. Non-metric MDS, using the rank order of distances rather than imputed actual distances, is particularly appropriate for much social network data and operates in very much the same way. Useful implementations of this approach are now available in the major software packages, and novel techniques such as multiple correspondence analysis are also beginning to become available.

The use of MDS and similar scaling or dimensional models provides a visual image of a network, but it also provides measures of distance, direction, and location that are not available within graph theory or algebraic approaches. This can allow a more rigorous embedding of social networks in social space and can permit intriguing hypotheses to be formed. For example, cliques and components, centrality and density, can all be mapped onto an MDS solution, and the relations among cliques or central points appear strikingly and can be measured determinately. The procedures made available within the *Pajek* program [18] have been designed specifically to allow such possibilities for large-scale networks. Graph theory and algebraic methods can be used together with the spring embedding of data through a variant of MDS to produce powerful network representations.

Particular problems of visualization occur when research is concerned with two-mode data rather than simple one-mode data. The most easily interpretable results are those that display a bipartite graph by using different symbols or colors for the two sets of data. Such graphs are a logical extension of graph theory approaches to two-mode data. Some interesting work is now being undertaken, however, using Galois lattices. Abandoning graph theoretical and spatial representations, the lattices aim to disclose the pattern of connections and their hierarchical order of "containment" from core to periphery.

One particularly striking advance in visualization techniques has been the use of models that allow the tracking of change over time. These have typically involved the adaptation of animation methods. Further advances in this area might be expected as some of the techniques discussed below begin to offer improved ways of handling temporal data on social networks.

Statistical Models for Hypothesis Testing

Much work on social networks has remained at the level of description and has failed to move towards explanatory concerns. Only rarely have researchers gone beyond a description to formulate hypotheses and test theories aimed at explaining the observed structures. One reason for this absence of theoretical work and hypothesis testing has been the weakness of the available statistical techniques for undertaking this work. While attempts have sometimes been made to use basic statistical measures of probability and significance to test hypotheses about network structure, these involve a number of difficulties. Standard statistical procedures such as significance tests, regression, and the analysis of variance all assume the independence of observations. The relational data studied by social network analysts, however, cannot be seen in this way. Relational data are, by definition, non-independent observations. Thus, novel statistical techniques have been required if methods of statistical inference are to be used. The most important work in recent years has been the generalization of Markov graphs to a larger family of models that has been undertaken by Stanley Wasserman and his colleagues [44,46,65], building on earlier work by Frank and Strauss [21]. Their $\boldsymbol{p}^*$ models – now termed exponential random graph models (ERG models or ERGM) – are designed specifically to allow the easier formulation and testing of theories. An ERGM defines a probability distribution on the set of all networks that can be constructed on a given set of points in terms of a particular parameter vector. ERGMs comprise a family of statistical models that promise a great advance in the understanding of social networks and have the potential to connect with the "small world" methods discussed below. The models are discussed in ► Social Networks, Exponential Random Graph ($\boldsymbol{p}^*$) Models for.

For any set of points it is possible to generate through simulation a set of all the possible graphs connecting them. These randomly generated graphs vary along the full range from completely unconnected to completely connected. A very large number of such graphs can exist along that range, as the number of possible graphs varies with the size of the network at an exponential rate. Thus, the number of possible configurations for a three-point graph is 64, that of a four-point graph is 1024, and the number for a five point graph is in excess of a million. An actually observed network can be treated as one realization from this set of logically possible graphs. If its probability of occurring by chance is low, then its actual occurrence may be regarded as statistically significant.

To assess this statistical significance, it is necessary to construct a probability distribution of the random graphs. Each random graph has a specific probability of occurring, depending on its particular structural properties. The aim in constructing these distributions is to see whether a particular structural property, such as a particular level of density, centralization, or clustering, is more or less likely to occur by chance than any other. These probabilities are estimated through log linear analysis from the statistical dependencies among the members of the network. The log odds ratios of the conditional probabilities of the relational ties associated with each element constructed from this dependence graph are used to produce an approximate likelihood function that can be used with Monte Carlo estimation techniques to produce maximum likelihood estimates. In this process the successive approximations constitute chains of graphs – a Markov chain – that converge towards a stationary distribution and a stable estimate of network parameters. Thus, ERGM provides a means for testing the probability that the observed network will occur with precisely the structural characteristics that it has on the basis of chance alone. This enables the researcher to highlight the statistically significant – non-chance – results apparent in the data.

Agent-Based Computational Models and Temporal Processes

As well as its descriptive focus, much social network analysis has also concentrated on the static features of social networks. This has also begun to change in recent years as more attention has been given to the dynamic processes involved in changes over time. A key advance in this direction has come from the use of models that depict the ways in which the behavior of individual agents results in global transformations of network structure. In so-called agent-based computational models, agents (whether individuals or groups) are seen as rule-following entities whose decisions to act in one way or another are consequential for the overall network by virtue of their concatenation with the action consequences of others. Therefore, a knowledge of the rules under which agents act can be used to predict broad patterns of change in network structure.

Agent-based models originated in simulation studies, where a simple and determinate set of rules can be applied to generate specific network structures (see also pp. 99ff in [25]; [40]). Applications of this idea to human social networks have tended to rely on assumptions drawn from rational choice theory, according to which individual agents are assumed to act on certain simple principles of utility maximization. However, the more complex models allow for the handling of different categories of agent, the members of each category following a different set of rules, and for a relaxation of the assumption of pure rationality by modeling rules of non-rational or ritualistic behavior. Such models are stochastic – specifying the probability of certain courses of action but not implying a complete determinism.

A particularly powerful example of such an approach is that of Tom Snijders [52,53,54] (see ▶ Network Analysis, Longitudinal Methods of). Change in network structure is regarded as a probabilistic process, describable by an objective function that measures the attractiveness to the agents of changes in their connectivity. The model sees an incremental adjustment of individual action to the changing network structure, resulting in a continuous – but often non-linear – process of network evolution. It is assumed that agents control the outgoing relations to others within the constraints on their opportunities that are set by the current network structure. However, they act "myopically" in relation to the immediate environment and the immediate consequences of their actions, having no conception of the wider consequences of their choices. At each point in time, the outcome for the overall network is a result of the intersection of the courses of action pursued by these myopic agents and so it will rarely correspond to the intentions of any one actor. At the same time, each actor has only a partial awareness of the changes that have resulted from their actions and so their next acts must be regarded as equally myopic and as likely, once again, to have unintended consequences at the network level. As a result of the continual iteration of actions, there is a co-evolution of network and behavior. Complex patterns of change are likely to occur. Snijders shows that small, incremental changes can accumulate to a point of "crisis" [55], a tipping point at which a radical, non-linear transformation of network structure occurs.

Snijders' longitudinal approach is especially powerful as it explicitly takes account of global network parameters such as density, reciprocity, distance, and balance. It can also be reprogrammed to recognize a wide range of different rule systems and decision procedures. Snijders also suggests that the approach can handle a greater degree of far-sightedness in agents, though with a corresponding decline in the determinateness of the analyses that can be drawn. Current work is making important connections with the early work of Wasserman [63] and the subsequent development of ERGM procedures. The approach is implemented in the *SIENA* program and has been used in a number of simulation studies.

Small World Models and Network Dynamics

The most recent area of development in social network analysis is the investigation of small world networks inspired by a number of physicists interested in network dynamics (explored in various entries in other sections of this *Encyclopedia*). Although this has, as yet, led to relatively few empirical investigations, there is growing interest in these applications. Many advocates of the approach have failed to discover or to appreciate the amount of work already undertaken in social network analysis and have assumed a very truncated history for the approach. There are, however, some signs of a closer appreciation of social science work, and the approaches are beginning to have an impact in the mainstream of social network analysis.

The "small world" hypothesis suggests that any two people chosen at random are likely to be connected to each other by a path of relatively short length. The hypothesis was formulated by Ithiel de Sola Pool and Manfred Kochen in a working paper of 1958 that was not published for another 20 years [17]. Their ideas achieved a wide currency and in 1967 Stanley Milgram at Harvard undertook some experimental studies to test the hypothesis [38]. He gave experimental subjects the names of people initially unknown to them and asked them to get a message to them using only known persons as intermediaries. That is, the messages were to pass through friend of friends of friends until they reached their destinations. Milgram discovered that subjects and targets were typically separated by paths of distance 6, no matter what the physical distances separating them. This finding embodied the common exclamation "it's a small world" uttered when people discover a mutual connection. Milgram hypothesized that the density of links of acquaintanceship is such that a considerable amount of redundancy is built into social networks: it is level of density in a network that explains its small world characteristics.

In the late 1990s, Watts and Strogatz [68] undertook some empirical studies in the mathematics of random networks. The structural properties of random networks, produced through the random linking of one point to another, had first been studied by Paul Erdös and Alfréd Rényi [20] and had led to a number of interesting conclusions. Watts and Strogatz discovered that a particular subset of random graphs corresponded closely to those actually found in natural and social phenomena, and they demonstrated that these had the small world characteristics identified by Milgram; though they were skeptical about the precise finding of "six degrees of separation". They showed, however, that the high levels of clique and sub-graph formation that were basic to the redundancy found in real networks gave them a number of characteristic features. Crucial to the structure of these networks was the existence of well-connected points – which they termed "hubs" – that make the density and, therefore, the small world properties, possible. It is the existence of hubs that explains why the density does not need to be extremely high for small world conditions to apply. On the contrary, there is a middle range of density in which efficient communication is possible.

The physicist Albértó-Laszlo Barabási, searching for a new topic to investigate, began to look at the findings of Watts and Strogatz. Barabási's central idea was that of the "power law" or "scale free" distribution, according to which small world networks have a frequency distribution of point degrees showing a small number of well-connected points and a large number of less well-connected points. These well-connected points are the hubs that generate the small world properties. He concluded that these findings provided the basis for a completely new approach to networks [1]. As has already been noted, his claims to novelty and to the revolutionary character of the work seriously underestimated the amount of prior work undertaken on social networks and their properties. Nevertheless, he did formulate some intriguing ideas and the work subsequently taken up by Watts does offer the prospect of genuine advances in social network analysis.

Watts holds that a small world graph is one in which there are a large number of "short cuts". These are lines that connect points that would otherwise be quite distant from each other. Neither a completely connected graph nor a sparse graph have short cuts, and so small world properties occur in an intermediate range of graphs. They are locally clustered but globally sparse. It is for these graphs, he holds, that precise mathematical conclusions about structure and structural development can be drawn. He holds, for example, that relatively small changes in network connectivity, such as the addition of a small number of shortcuts, can dramatically change its properties through "phase transitions" at critical threshold points. There are considerable possibilities inherent in using these ideas alongside those from agent-based computational studies already discussed. The common thread is a concern for the investigation of longitudinal processes in which micro-level incremental changes result in non-linear macro-level transformations of structure.

Bibliography

1. Barabási A-L (2002) Linked: The new science of networks. Perseus, Cambridge
2. Barnes JA (1954) Class and committee in a Norwegian island parish. Hum Relat 7:39–58
3. Bavelas A (1948) A mathematical model for group structures. Appl Anthr 7:16–30
4. Berkowitz SD (1982) An introduction to structural analysis. Butterworths, Toronto
5. Boorman SA, White HC (1976) Social structure from multiple networks, vol II. Am J Sociol 81:1384–1446
6. Bott E (1955) Urban families: Conjugal roles and social networks. Hum Relat 8:253–292
7. Bott E (1956) Urban families: The norms of conjugal roles. Hum Relat 9:325–342
8. Buchanan M (2002) Small world: Uncovering nature's hidden networks. Weidenfeld and Nicolson, London
9. Burt RS (1980) Models of network structure. Ann Rev Sociol 6:79–141
10. Burt RS (1982) Towards a structural theory of action. Academic Press, New York
11. Burt RS (1992) Structural holes. Cambridge University Press, New York
12. Carrington PJ, Scott J, Wasserman S (2005) Models and methods in social network analysis. Cambridge University Press, Cambridge
13. Cartwright D, Zander A (1953) Group dynamics. Tavistock, London
14. Coleman JS, Katz E, Menzel H (1966) Medical innovation: A diffusion study. Bobbs-Merrill, New York
15. Davis AB, Gardner BB, Gardner MR (1941) Deep south. University of Chicago Press, Chicago
16. De Nooy W, Mrvar A, Batagelj V (2005) Exploratory social network analysis with Pajek. Cambridge University Press, New York
17. De Sola Pool I, Kochen M (1978) Contacts and influence. Soc Netw 1(1):5–51
18. DeNooy W, Mrvar A, Batagelj V (2005) Exploratory social network analysis with Pajek. Cambridge University Press, New York
19. Doreian P, Batagelj V, Ferligoj A (2005) Generalized blockmodelling. Cambridge University Press, New York
20. Erdös P, Rényi A (1959) On the evolution of random graphs. Publ Math Inst Hung Acad Sci 5:17–61
21. Frank O, Strauss D (1986) Markov Graphs. J Am Stat Assoc 81:832–842
22. Freeman LC (1979) Centrality in social networks: Conceptual clarification. Soc Netw 1:215–239
23. Freeman LC (2004) The development of social network analysis: A study in the sociology of science. Empirical Press, Vancouver
24. Freeman LC, White DR, Romney AK (1989) Research methods in social network analysis. Transaction Books, New Brunswick
25. Gilbert N, Troizch KG (1999) Simulation for the social scientist. Open University Press, Buckingham
26. Granovetter M (1973) The strength of weak ties. Am J Sociol 78:1360–1380
27. Granovetter M (1974) Getting a job. Harvard University Press, Cambridge
28. Harary F, Norman RZ (1953) Graph theory as a mathematical model in social science. Institute for Social Research, Ann Arbor
29. Helmers HM (1975) Graven naar macht. Van Gennep, Amsterdam
30. Homans G (1950) The human group. Routledge and Kegan Paul, London
31. Homans G (1961) Social behaviour: Its elementary forms. Routledge and Kegan Paul, London
32. Knoke D, Kuklinski JH (1982) Network analysis. Sage Publications, Beverley Hills
33. König D (1936) Theorie der endlichen und unendlichen Graphen. Chelsea, New York
34. Laumann EO (1966) Prestige and association in an urban community. Bobbs-Merrill, Indianapolis
35. Levine JH (1972) The sphere of influence. Am Sociol Rev 37:14–27
36. Lorrain F, White HC (1971) Structural equivalence of individuals in social networks. J Math Sociol 1:49–80
37. Marsden PV, Lin N (1982) Social structure and network analysis. Sage, Beverley Hills
38. Milgram S (1967) The small world problem. Psychol Today 2:60–67
39. Mitchell JC (1969) Social networks in urban situations. Manchester University Press, Manchester
40. Monge PR, Contractor NS (2003) Theories of communication networks. Oxford University Press, Oxford
41. Moreno JL (1934) Who shall survive? Beacon Press, New York
42. Nadel SF (1957) The theory of social structure. Free Press, Glencoe
43. Pattison P (1993) Algebraic models for social networks. Cambridge University Press, Cambridge
44. Pattison P, Wasserman S (1999) Logit models and logistic regressions for social networks: II Multivariate relations. Br J Math Stat Psychol 52:169–193
45. Radcliffe-Brown AR (1940) On social structure. In: Radcliffe-Brown AR (ed) Structure and function in primitive society. Cohen and West, London
46. Robins GL, Pattison P, Wasserman S (1999) Logit models and logistic regressions for social networks: III Valued relations. Psychometrika 64:371–394
47. Roethlisberger FJ, Dickson WJ (1939) Management and the worker. Harvard University Press, Cambridge
48. Rogers E (1962) Diffusion of innovations, 5th edn. Free Press, New York
49. Ryan R, Gross N (1943) The diffusion of hybrid seed corn in two Iowa communities. Rural Sociol 8:15–24
50. Scott J (1991) Networks of corporate power: A comparative assessment. Ann Rev Sociol 17:181–203
51. Scott J (2000) Social network analysis, 2nd edn. Sage, London
52. Snijders TAB (2001) The statistical evaluation of social network dynamics. In: Sobel ME, Becker MP (eds) Sociological methodology. Basil Blackwell, Oxford
53. Snijders TAB (2005) Models for longitudinal network data. In: Carrington PJ, Scott J, Wasserman S (eds) Models and methods in social network analysis. Cambridge University Press, Cambridge

54. Snijders TAB, Van Duijn MAJ (1997) Simulation for statistical inference in dynamic network models. In: Conte R, Hegelmann R, Terna P (eds) Simulating social phenomena. Springer, Berlin
55. Thom R (1972) Structural stability and morphogenesis. Benjamin, London
56. Travers J, Milgram S (1969) An experimental study of the small world problem. Sociometry 32:425–443
57. Valente TW (1995) Network models of the diffusion of innovations. Hampton Press, Cresskill
58. Warner WL (1949) Democracy in Jonesville. Harper and Brothers, New York
59. Warner WL, Low JO (1947) The social system of a modern factory. Yale University Press, New Haven
60. Warner WL, Lunt PS (1941) The social life of a modern community. Yale University Press, New Haven
61. Warner WL, Lunt PS (1942) The status system of a modern community. Yale University Press, New Haven
62. Warner WL, Srole L (1945) The social systems of american ethnic groups. Yale University Press, New Haven
63. Wasserman S (1980) Analyzing social networks as stochastic processes. J Am Stat Assoc 75:280–294
64. Wasserman S, Faust K (1994) Social network analysis: Methods and applications. Cambridge University Press, New York
65. Wasserman S, Pattison P (1996) Logit models and logistic regressions for social networks, vol I. An introduction to Markov random graphs and $\boldsymbol{p}^*$. Psychometrika 60:401–426
66. Watts D (1999) Small worlds: The dynamics of networks between order and randomness. Princeton University Press, Princeton
67. Watts D (2003) Six degrees. The science of a connected age. Norton WW, New York
68. Watts DJ, Strogatz SH (1998) Collective dynamics of small-world networks. Nature 393:440–442
69. Wellman B (1979) The community question: The intimate networks of East Yorkers. Am J Sociol 84:1201–1231
70. Wellman B (1988) Networks as personal communities. In: Wellman B, Berkowitz SD (eds) Social structures. Cambridge University Press, Cambridge
71. Wellman B, Berkowitz S (1988) Social structures. Cambridge University Press, New York
72. White HC (1963) An anatomy of kinship. Prentice-Hall, Englewood Cliffs
73. White HC, Boorman SA, Breiger RL (1976) Social structure from multiple networks, I. Am J Sociol 81:1384–1446

Social Network Analysis, Two-Mode Concepts in

STEPHEN P. BORGATTI
Gatton College of Business and Economics,
University of Kentucky, Lexington, USA

Article Outline

Glossary

2-Mode matrix A (2-dimensional) matrix is said to be 2-mode if the rows and columns index different sets of entities (e. g., the rows might correspond to persons while the columns correspond to organizations). In contrast, a matrix is 1-mode if the rows and columns refer to the same set of entities, such as a city-by-city matrix of distances.

Blockmodel A blockmodel is a partitioning of the cells of a matrix into blocks that is induced by the partitioning the rows and columns into classes and sorting the matrix such that rows (and columns) that belong to the same class are next to each other. More specifically, two matrix cells x_{ij} and x_{mp} are in the same block if $\text{class}(i) = \text{class}(m)$ and $\text{class}(j) = \text{class}(p)$.

Centrality A family of concepts for characterizing the structural importance of a node's position in a network.

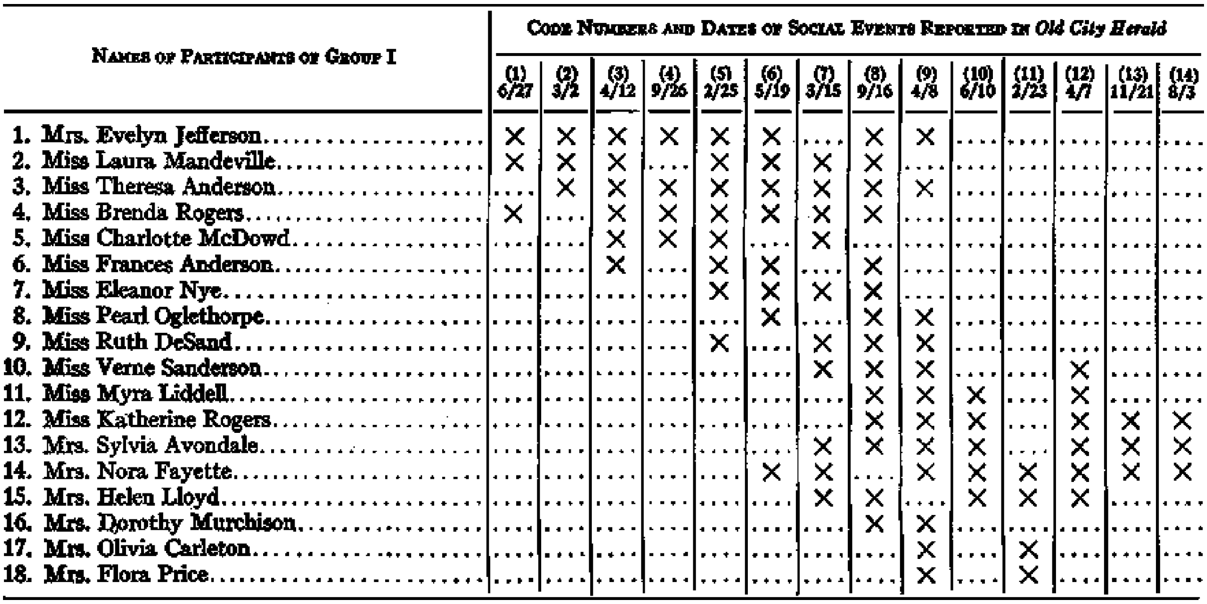
Graph cohesion A family of concepts characterizing the extent of connectedness of a graph, such as density (the proportion of pairs of nodes that have ties with each other), or average path distance.

Multidimensional scaling (MDS) A method of locating points in space such that Euclidean distances between the points correspond to a matrix of input similarities/distances among objects. Used to provide visual representations of 1-mode matrices such as correlation matrices or perceptual distances among objects.

Regular equivalence The definition of regular equivalence is recursive. If two nodes are regularly equivalent, then they are connected to regularly equivalent nodes. Regular equivalence is used to identify nodes that are playing the same structural role, even if they are not connected to each other.

Social network (or, in graph theory, a graph) A collection of nodes (also referred to as vertices or actors) together with a set of ties (also known as edges or links) that connect pairs of nodes. Typically used to represent social relations such as who is friends with whom, or who is the supervisor of whom.

Structural equivalence At an intuitive level, a pair of nodes is said to be structurally equivalent to the ex-



Names of Participants of Group I	Code Numbers and Dates of Social Events Reported in *Old City Herald*													
	(1) 6/27	(2) 3/2	(3) 4/12	(4) 9/26	(5) 2/25	(6) 5/19	(7) 3/15	(8) 9/16	(9) 4/8	(10) 6/10	(11) 2/23	(12) 4/7	(13) 11/21	(14) 8/3
1. Mrs. Evelyn Jefferson	×	×	×	×	×	×	....	×	×	....	....	....	....	....
2. Miss Laura Mandeville	×	×	×	....	×	×	×	×	....	....	....	....	....	....
3. Miss Theresa Anderson	....	×	×	×	×	×	×	×	×	....	....	....	....	....
4. Miss Brenda Rogers	×	....	×	×	×	×	×	×	....	....	....	....	....	....
5. Miss Charlotte McDowd	....	....	×	×	×	....	×	....	....	....	....	....	....	....
6. Miss Frances Anderson	....	....	×	....	×	×	....	×	....	....	....	....	....	....
7. Miss Eleanor Nye	....	....	....	....	×	×	×	×	....	....	....	....	....	....
8. Miss Pearl Oglethorpe	....	....	....	....	....	×	....	×	×	....	....	....	....	....
9. Miss Ruth DeSand	....	....	....	....	×	....	×	×	×	....	....	....	....	....
10. Miss Verne Sanderson	....	....	....	....	....	....	×	×	×	....	....	×	....	....
11. Miss Myra Liddell	....	....	....	....	....	....	....	×	×	×	....	×	....	....
12. Miss Katherine Rogers	....	....	....	....	....	....	....	×	×	×	....	×	×	×
13. Mrs. Sylvia Avondale	....	....	....	....	....	....	×	×	×	×	....	×	×	×
14. Mrs. Nora Fayette	....	....	....	....	....	×	×	....	×	×	×	×	×	×
15. Mrs. Helen Lloyd	....	....	....	....	....	....	×	×	....	×	×	×	....	....
16. Mrs. Dorothy Murchison	....	....	....	....	....	....	....	×	×	....	....	....	....	....
17. Mrs. Olivia Carleton	....	....	....	....	....	....	....	....	×	....	×	....	....	....
18. Mrs. Flora Price	....	....	....	....	....	....	....	....	×	....	×	....	....	....

Social Network Analysis, Two-Mode Concepts in, Figure 1
DGG women-by-events matrix

tent that they occupy identical locations in a network, meaning that they are connected to exactly the same others. Structurally equivalent nodes are identical with respect to all structural properties, such as centrality or subgroup membership.

Definition of the Subject

In social network analysis, 2-mode data refers to data recording ties between two sets of entities. In this context, the term "mode" refers to a class of entities – typically called actors, nodes or vertices – whose members have social ties with other members (in the 1-mode case) or with members of another class (in the 2-mode case). Most social network analysis is concerned with the 1-mode case, as in the analysis of friendship ties among a set of school children or advice-giving relations within an organization. The 2-mode case arises when researchers collect relations between classes of actors, such as persons and organizations, or persons and events. For example, a researcher might collect data on which students in a university belong to which campus organizations, or which employees in an organization participate in which electronic discussion forums. These kinds of data are often referred to as affiliations. Co-memberships in organizations or participation in events are typically thought of as providing opportunities for social relationships among individuals (and also as the consequences of pre-existing relationships). At the same time, ties between organizations through their members are thought to be conduits through which organizations influence each other.

Introduction

Perhaps the best known example of 2-mode network analysis is contained in the study of class and race by Davis, Gardner and Gardner (henceforth DGG) published in the 1941 book *Deep South* [6]. They followed 18 women over a nine-month period, and reported their participation in 14 events, such as a meeting of a social club, a church event, a party, and so on. Their original figure is shown in Fig. 1.

DGG used the data to investigate the extent to which social relations tended to occur within social classes.

Basic Concepts

A typical data matrix has two dimensions or *ways*, corresponding to the rows and columns of the matrix. The number of ways in a matrix X can be thought of as the number of subscripts needed to represent a particular datum, as in x_{ij}. If we stack together a number of similarly sized 2-dimensional matrices, we can think of the result as a 3-dimensional or 3-way matrix.

The *modes* of a matrix correspond to the distinct sets of entities indexed by the ways. In the DGG dataset described above, the rows correspond to women and the columns to a different class of entities, namely events.

Hence, the matrix has two modes in addition to two ways; it is 2-way, 2-mode. In contrast, a persons-by-persons matrix A, in which $a_{ij} = 1$ if person i is friends with person j, is a 2-way, 1-mode matrix, because both ways point to the same set of entities.

In a sense, what constitutes different modes is up to the researcher. If we collect romantic ties among a group of heterosexual people of both genders, we could construct a 2-mode men-by-women matrix X in which $x_{ij} = 1$ if a romantic tie was observed between man i and woman j, and $x_{ij} = 0$ otherwise. Or, one could construct a larger 1-mode person-by-person matrix B also consisting of 1s and 0s in which it just happens that 1s only occur in cells where the row and column correspond to persons of different gender. Use of the men-by-women matrix would imply that same-gender relations were impossible, whereas use of the person-by-person matrix would suggest that same-gender relations were logically possible, even if actually not observed.

Matrices recording relational information such as romantic ties can be represented as mathematical graphs as well. A graph $G(V, E)$ consists of a set of nodes or vertices V together with a set of lines or edges E that connect them. An edge is simply an unordered pair of nodes (u, v). (In directed graphs or *digraphs* we use *ordered* pairs to indicate direction of the tie.) To indicate a tie between two nodes u and v, we simply include the pair (u, v) in the set E. The number of nodes in a graph is denoted by $|V|$ or n.

A bipartite graph is a graph in which we can partition all nodes into two sets, V_1 and V_2, such that all edges include a member of V_1 and a member of V_2. The number of nodes in each vertex set is denoted n_1 and n_2, respectively.

Two-Mode Data in Social Network Analysis

Most social networks are conceived of as relations among a set of nodes, and therefore represented as a 1-mode matrix (typically of 1s and 0s) or a simple graph or digraph. For example, we might collect data on who is friendly with whom within an organization, or who injects drugs with whom in a neighborhood.

However, 2-mode data are common in social network contexts as well. Typical examples include, actor-by-event attendance (as in the DGG data), actor by group membership (such as managers sitting on corporate boards), and actor by trait possession (such as adjective checklist data), and actor by object possession (such as material style of life scales in which inventories are made of household possessions).

In many cases when 2-mode data are collected, the analytical interest is focused on one mode or the other. For example, in the DGG dataset, person-by-event attendances were collected in order to understand social relations among the women, specifically, whether women tended to have social relations primarily within their own social classes. In the interlocking directorate literature, membership of executives on corporate boards is collected mainly in order to understand how corporations are intertwined, and how the structure of this connectivity affects corporate control of society.

However, it can also occur that neither mode dominates our analytical focus and the primary interest is in the correspondence of one mode to the other. For example, a university might ask its faculty which courses they prefer to teach. Here, the objective is typically not to understand how faculty are related to each other through courses, nor how courses are related via faculty, but in the optimal assignments of persons to courses so that courses are staffed and faculty are not complaining.

Unimodal Approaches to Two-Mode Data

One approach to handling 2-mode data in social network analysis is to convert the data to 1-mode data. This is especially appropriate when the analytical interest focuses primarily on just one of the modes. Consider, for example, the case of a person by group membership matrix X in which $x_{ij} = 1$ if person i belongs to group j. Let us assume that the groups are small and everyone in a group knows everyone else. In that case, we could try to infer an acquaintance network by constructing a 1-mode matrix A such that $a_{ij} = 1$ if person i is in at least one group with person j. Better yet, we can construct a valued matrix A such that a_{ij} gives the number of groups that i and j are both members of. In other words,

$$a_{ij} = \sum_k x_{ik} x_{jk} \text{ or } A = \mathrm{XX}' \, . \tag{1}$$

We might regard a_{ij} as a proxy for the social proximity of i and j, or perhaps as a rough indicator of the potential for information flow between them. In this approach, we analyze each mode of the data separately. Figure 2 shows the values of A for the 2-mode data shown in Fig. 1.

It should be noted that A can be seen as a matrix of profile similarities or correlations among pairs of rows in X. For example, the matrix of Pearson correlations among rows of X is defined as follows:

$$r_{ij} = \frac{\frac{1}{m} \sum_k x_{ik} x_{jk} - u_i u_j}{s_i s_j} \, , \tag{2}$$

where u_i is the mean of row i and s_i is the standard deviation of row i. It is evident that the correlation r_{ij} is es-

	EVE	LAU	THE	BRE	CHA	FRA	ELE	PEA	RUT	VER	MYR	KAT	SYL	NOR	HEL	DOR	OLI	FLO
EVELYN	8	6	7	6	3	4	3	3	3	2	2	2	2	2	1	2	1	1
LAURA	6	7	6	6	3	4	4	2	3	2	1	1	2	2	2	1	0	0
THERESA	7	6	8	6	4	4	4	3	4	3	2	2	3	3	2	2	1	1
BRENDA	6	6	6	7	4	4	4	2	3	2	1	1	2	2	2	1	0	0
CHARLOTTE	3	3	4	4	4	2	2	0	2	1	0	0	1	1	1	0	0	0
FRANCES	4	4	4	4	2	4	3	2	2	1	1	1	1	1	1	1	0	0
ELEANOR	3	4	4	4	2	3	4	2	3	2	1	1	2	2	2	1	0	0
PEARL	3	2	3	2	0	2	2	3	2	2	2	2	2	2	1	2	1	1
RUTH	3	3	4	3	2	2	3	2	4	3	2	2	3	2	2	2	1	1
VERNE	2	2	3	2	1	1	2	2	3	4	3	3	4	3	3	2	1	1
MYRNA	2	1	2	1	0	1	1	2	2	3	4	4	4	3	3	2	1	1
KATHERINE	2	1	2	1	0	1	1	2	2	3	4	6	6	5	3	2	1	1
SYLVIA	2	2	3	2	1	1	2	2	3	4	4	6	7	6	4	2	1	1
NORA	2	2	3	2	1	1	2	2	2	3	3	5	6	8	4	1	2	2
HELEN	1	2	2	2	1	1	2	1	2	3	3	3	4	4	5	1	1	1
DOROTHY	2	1	2	1	0	1	1	2	2	2	2	2	2	1	1	2	1	1
OLIVIA	1	0	1	0	0	0	0	1	1	1	1	1	1	2	1	1	2	2
FLORA	1	0	1	0	0	0	0	1	1	1	1	1	1	2	1	1	2	2

Social Network Analysis, Two-Mode Concepts in, Figure 2
Women-by-women matrix of overlaps across events

		Row *j*	
		1	0
Row *i*	1	*a*	*b*
	0	*c*	*d*

Social Network Analysis, Two-Mode Concepts in, Figure 3
Contingency table *T*

sentially a_{ij} corrected for the number of groups that each belongs to. This kind of correction seems eminently desirable, but of course there are many ways of doing this. For example, consider a cross-tabulation T of row i and row j, such that t_{uv} gives the number of columns k of X for which $x_{ik} = u$ and $x_{jk} = v$, as shown in Fig 3.

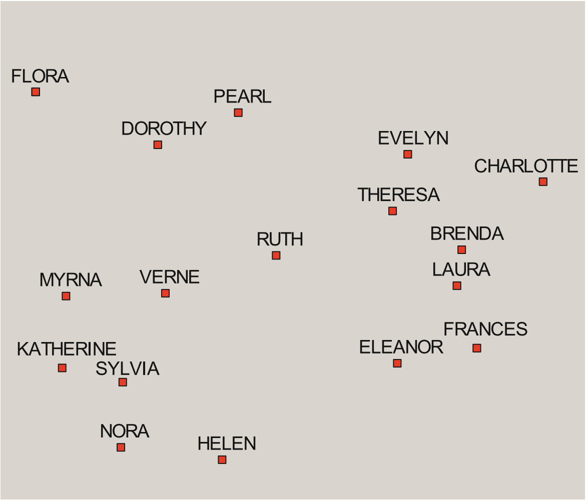
In the table, the first row marginal $(a + b)$ gives the number of groups that person i belongs to while the first column marginal gives the number of groups that person j belongs to. Note that a_{ij} of Eq. (1) corresponds to a in Fig. 3. An obvious approach is the Jaccard coefficient, which may be defined as

$$c_{ij} = \frac{a}{a+b+c} \,. \tag{3}$$

Thus, c_{ij} is essentially the cardinality of the intersection of the groups belonged to by both persons, divided by the cardinality of their union. The Jaccard coefficient is often recommended when the number of columns in X is large and the number of 1s in each row is highly limited.

An alternative specifically designed for 2-mode affiliation data is given by Bonacich [1]. It normalizes a_{ij} as follows:

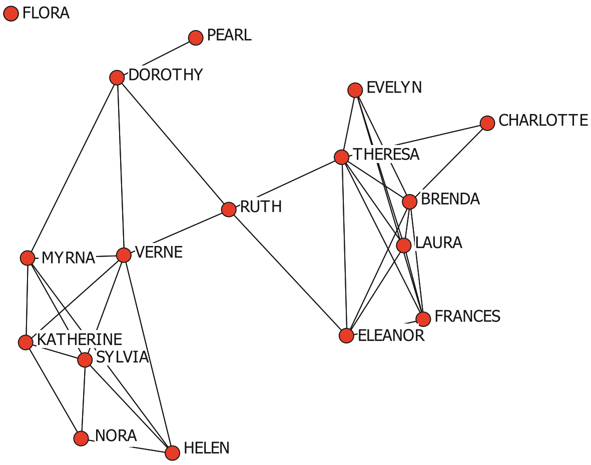
$$a^*_{ij} = \frac{a_{ij} - \sqrt{adbc}}{ad - bc} \,, \quad \text{for } ad \neq bc \,. \tag{4}$$



Social Network Analysis, Two-Mode Concepts in, Figure 4
Multidimensional scaling of Jaccard coefficients among rows of DGG matrix

Unimodal Visualization of Two-Mode Data

Given that a 2-mode matrix has been transformed into a 1-mode matrix by taking similarities among the rows (or columns), one can visualize the network using all the usual techniques for visualization of valued networks. For exam-



Social Network Analysis, Two-Mode Concepts in, Figure 5
Multidimensional scaling of Jaccard similarity coefficients, with lines indicating similarity scores greater than 0.4. Actor Olivia is obscured by Flora

ple, a standard approach is to use metric multidimensional scaling (MDS) on the matrix A, generating a map in which points corresponding to nodes (e. g., persons) appear close to each other to the extent that they share many groups. Figure 4 shows such an MDS map based on Jaccard similarities.

To highlight structure, it is common to overlay lines between pairs of nodes with a similarity score greater than a certain level. Figure 5 shows the result for similarities greater than 0.4. The diagram effectively shows the bridging role of Ruth, who, based on ethnographic evidence, was seen by DGG as a member of both groups of women.

Alternatively, one can use a standard graph layout algorithm (GLA) to draw the graph induced by dichotomizing the Jaccard similarity matrix. For example, define $(u, v) \in E$ if and only if (iff) $c_{ij} > 0.4$. Compared to multidimensional scaling representations, GLAs have the disadvantage that distances between points cannot strictly be interpreted, but this property also means that nodes need not obscure each other. Figure 6 shows the results of applying a spring-embedding [10] GLA to the dichotomized data.

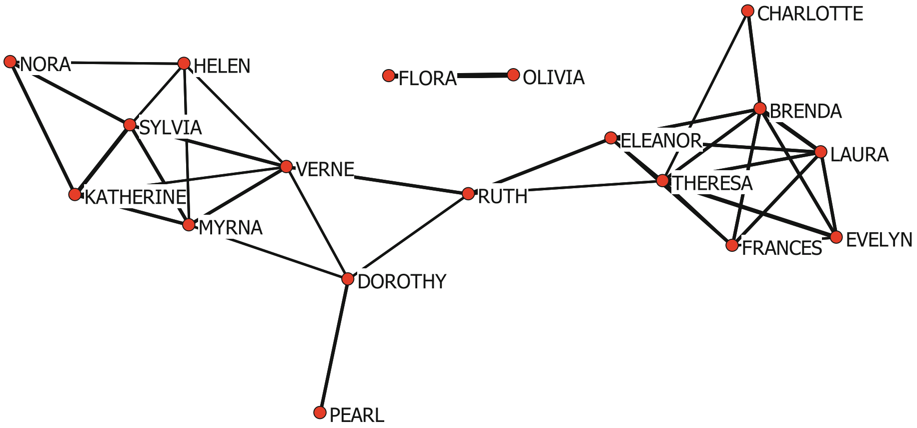
A similar analysis can be carried out on the events rather than the women. Applying Eq. (3) to the columns of the 2-mode matrix in Fig. 1 yields a matrix of Jaccard coefficients which can be visualized using the same methods used for the women. Figure 7 shows events with Jaccard overlaps greater than 0.35.

Unimodal Analysis of Two-Mode Data

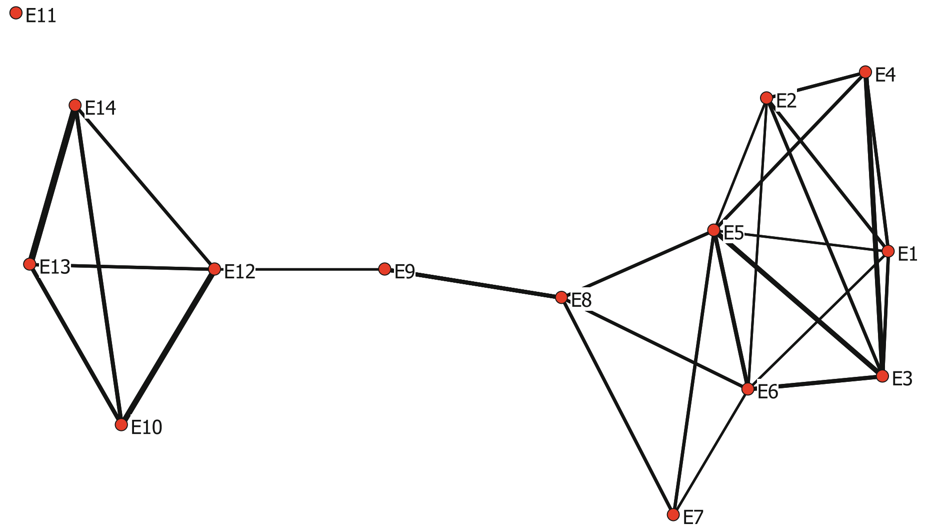
In general, analysis of 2-mode data transformed into valued 1-mode networks proceeds like any other valued network. As with visualization, this often means generating a graph from the valued data via some rule such as $(u, v) \in E$ iff $a_{ij} > q$, where q is chosen by the researcher. Typically, there is no theoretical reason for choosing any particular value of q; hence a series of different values is generally chosen and the analysis repeated for each.

There are, however, a few consequences that stem from the 2-mode origin of the data. By their very nature, many commonly used measures of similarity and dissimilarity satisfy triangle inequality laws. For example, for Euclidean distance, every triple of nodes i, j, k satisfy the following rule:

$$d_{ik} \leq d_{ij} + d_{jk} \,. \tag{5}$$



Social Network Analysis, Two-Mode Concepts in, Figure 6
Spring-embedding representation of Jaccard similarities dichotomized at > 0.4



Social Network Analysis, Two-Mode Concepts in, Figure 7
Spring-embedding representation of "ties" among events ($c_{ij} > 0.35$)

As a result, the 1-mode data (especially if not dichotomized) artifactually exhibit a certain level of transitivity that may be higher than baseline models built on simple sociometric choice data would expect. Statistics based on transitivity, such as structural holes and clustering coefficients, must similarly be interpreted with some caution in such data.

Bimodal Approaches to Two-Mode Data

Another approach to working with 2-mode data seeks to analyze both modes simultaneously. The data are seen to represent relations between two sets of nodes, forming a bipartite graph $G_B(V_1 + V_2, E)$ in which, or all u and v, $(u, v) \in E$ if and only if u and v belong to different vertex sets. In other words, all ties are between vertex sets and none are within-group. The matrix representation of such a graph can be a rectangular incidence matrix X (as in Fig. 1) or a square bipartite adjacency matrix B with $n = n_1 + n_2$ rows representing both modes, and an equal number of columns, also representing both modes. In the latter case, the original matrix X forms a submatrix of the larger adjacency matrix B in which both rows and columns index the $V_1 + V_2$ entities. The matrix B is composed of four blocks, two of which are empty, as shown in Fig. 8 Note that the original matrix X forms the top right quadrant of B, and its transpose forms the bottom left quadrant.

Bimodal Visualization of Two-Mode Data

All of the standard ways to visualize networks, such as MDS and GLAs, apply to bipartite graphs. For example, Fig. 9 shows a spring-embedding layout of the bipartite graph represented by the matrix in Fig. 8.

In the representation, two nodes are near each other roughly to the extent that the geodesic distance between them is short. Thus, events are near each other if they are attended by the same women (distance 2), and women are near each other if they attend the same events. In this example, the representation makes clear that there is a set of women on the left (Mryna, Helen, Katherine, Nora, Silvia, etc.) that attend a set of events exclusive to them (events 10 through 13), and another set of women (Evelyn, Theresa, Laura, Brenda, etc.) that have their own events (E1 through E4), and finally a set of events that both "circles" of women attend (events E6 through E9).

For small datasets, this bimodal visualization is often extremely effective for transmitting a holistic understanding of the whole dataset.

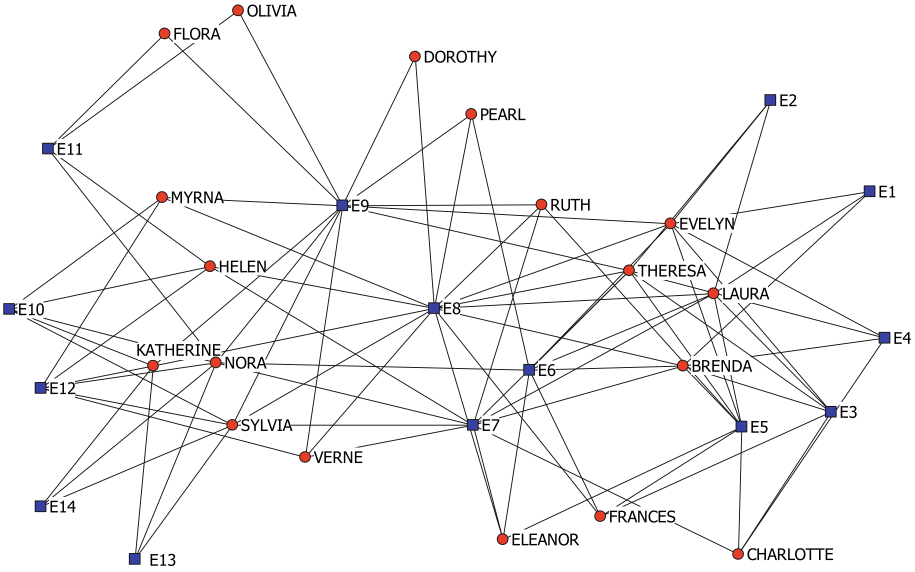
It is worth noting that there is a simple mathematical relationship between pairwise overlaps a_{ij} as defined in Eq. (1) and paths in the bipartite graph. Specifically, the number of 2-step paths between any pair of women i and j in the bipartite graph is equal to a_{ij}, the number of events they attended in common. Of course, the number of 2-step paths is simply the matrix product BB, the bipartite adjacency matrix multiplied by itself. As shown in Fig. 10, the top left block and bottom right blocks of BB are the matrices A as calculated by Eq. (1) for rows and columns respectively of X.

Bimodal Analysis of Two-Mode Data

Given a bimodal perspective, one approach to analyzing 2-mode data is to develop entirely new metrics and algo-

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	1	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1	1	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	0	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	1	1	1	0	1	0	1	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	1	1	1	1	1	1	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Social Network Analysis, Two-Mode Concepts in, Figure 8
Bipartite adjacency matrix *B* created from the original DGG 2-mode matrix *X*



Social Network Analysis, Two-Mode Concepts in, Figure 9
Spring-embedding representation of bipartite graph

rithms designed specifically for 2-mode data. Such techniques take cognizance of the fact that the observed network is not just bipartite by happenstance, but could not have been any other way. In other words, taking account of the fact that the observed lack ties between certain nodes (namely, those belonging to different modes) was by design – similar to the concept of structural zeros in log-linear modeling. To date, few techniques of this kind have

	EVE	LAU	THE	BRE	CHA	FRA	ELE	PEA	RUT	VER	MYR	KAT	SYL	NOR	HEL	DOR	OLI	FLO	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14
EVELYN	8	6	7	6	3	4	3	3	3	2	2	2	2	2	1	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LAURA	6	7	6	6	3	4	4	2	3	2	1	1	2	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
THERESA	7	6	8	6	4	4	4	3	4	3	2	2	3	3	2	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BRENDA	6	6	6	7	4	4	4	2	3	2	1	1	2	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CHARLOTTE	3	3	4	4	4	2	2	0	2	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
FRANCES	4	4	4	4	2	4	3	2	2	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ELEANOR	3	4	4	4	2	3	4	2	3	2	1	1	2	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PEARL	3	2	3	2	0	2	2	3	2	2	2	2	2	2	1	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RUTH	3	3	4	3	2	2	3	2	4	3	2	2	3	2	2	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VERNE	2	2	3	2	1	1	2	2	3	4	3	3	4	3	3	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MYRNA	2	1	2	1	0	1	1	2	2	3	4	4	4	3	3	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KATHERINE	2	1	2	1	0	1	1	2	2	3	4	6	6	5	3	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SYLVIA	2	2	3	2	1	1	2	2	3	4	4	6	7	6	4	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NORA	2	2	3	2	1	1	2	2	2	3	3	5	6	8	4	1	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HELEN	1	2	2	2	1	1	2	1	2	3	3	3	4	4	5	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DOROTHY	2	1	2	1	0	1	1	2	2	2	2	2	2	1	1	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OLIVIA	1	0	1	0	0	0	0	1	1	1	1	1	1	2	1	1	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
FLORA	1	0	1	0	0	0	0	1	1	1	1	1	1	2	1	1	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	2	3	2	3	3	2	3	1	0	0	0	0	0
E2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	3	3	2	3	3	2	3	2	0	0	0	0	0
E3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3	6	4	6	5	4	5	2	0	0	0	0	0
E4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	4	4	4	3	3	3	2	0	0	0	0	0
E5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3	6	4	8	6	6	7	3	0	0	0	0	0
E6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3	5	3	6	8	5	7	4	1	1	1	1	1
E7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	4	3	6	5	10	8	5	3	2	4	2	2
E8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3	5	3	7	7	8	14	9	4	1	5	2	2
E9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	2	2	3	4	5	9	12	4	3	5	3	3
E10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	4	4	5	2	5	3	3
E11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	1	3	2	4	2	1	1
E12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	4	5	5	5	2	6	3	3
E13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	2	3	3	1	3	3	3
E14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	2	3	3	1	3	3	3

Social Network Analysis, Two-Mode Concepts in, Figure 10
Matrix *BB* giving the number of paths of length 2 between all pairs of nodes

been developed, the exception being the area of 2-mode centrality measures, which has received significant attention (e. g., [2]).

Another approach is to treat the bipartite graphs as ordinary graphs and apply all the standard algorithms and techniques of social network analysis. Effectively this assumes either that the special nature of the graphs will not affect the techniques, or that we can pretend that ties within modes could have occurred and just didn't. This approach works for a small class of methods, but by no means all. For example, calculating transitivity fails because transitive triples are impossible in bipartite graphs (all ties are between modes, which means that if $a \rightarrow b$ and $b \rightarrow c$ then a and c must be members of the same class, and therefore cannot be tied, making transitivity impossible). In contrast, if we were to adapt the definition of transitivity to be based on quadruples such that a quad is transitive if $a \rightarrow b$, $b \rightarrow c$, $c \rightarrow d$ and $a \rightarrow d$, this would be an example of the first approach.

Finally, a compromise approach is to use the standard metrics and algorithms that apply to general graphs, and then either adjust the outputs via normalization or adjust the baseline expectations for the results [5,9]. In the former case, we develop normalizations that adjust the metrics (typically by dividing by theoretical bipartite maxima), and in the latter case we derive different theoretical distributions for the statistics in question. In general, we choose the former approach when statistics are bounded between 0 and 1 and can be interpreted as proportions of maximum possible values. We choose the latter approach when the statistics are unbounded and have direct interpretations (see example of cohesion below).

Bimodal Approaches to Graph Cohesion The simplest and most common measure of network cohesion is *density* – the number of edges divided by the number of pairs of nodes (using ordered pairs in the case of directed graphs and unordered pairs in the case of undirected graphs). In bipartite graphs, of course, only edges between vertex sets are possible. As a result, the maximum possible undirected ordinary density is

$$\frac{n_1 n_2}{(n_1 + n_2)(n_1 + n_2 - 1)} \,. \tag{6}$$

Thus, if density were calculated on a 2-mode network as if it were an ordinary graph, we would probably want to normalize the result by dividing by the sum above, otherwise the calculated density would appear misleadingly low. This is a case where the "compromise approach" discussed above is effective. (Of course, in terms of computation, it would be easier to simply calculate the average of the 2-mode matrix X rather than convert to bipartite form and then perform this adjustment.)

For measures of cohesion that are not expressed as fractions of maximum possible values, such as average geodesic path length, the need to renormalize is not as great, since the raw values are directly interpretable. In doing so, however, we must be careful not to compare the results with standard rules of thumb or theoretical models based on simple random graphs, as these have not assumed the bipartite restrictions.

Bimodal Approaches to Centrality Four measures of centrality are used with high frequency in social network analysis today: degree, closeness, betweenness and eigenvector. We discuss each in turn.

Degree centrality, d_i, is defined as the number of ties incident upon node i. Degree centrality is typically normalized by dividing by the maximum number of ties possible, which in a graph of n nodes is $n-1$. Hence

$$d_i^* = \frac{d_i}{n-1} \,. \tag{7}$$

However, in any (non-trivial) bipartite graph, no node can be connected to all others, since this would mean within-mode ties. Instead, for a node in V_1, the maximum number of ties it could have is n_2, whereas for a node in V_2, the maximum number of ties is n_1. Hence a natural adjustment is to provide two separate normalization formulas as follows:

$$\begin{aligned} d_i^* &= \frac{d_i}{n_2}\,, \quad \text{for } i \in V_1\,,\\ d_j^* &= \frac{d_j}{n_1}\,, \quad \text{for } j \in V_2\,. \end{aligned} \tag{8}$$

Closeness centrality is ordinarily defined as the sum of geodesic distances from a node to all others, as shown in Eq. (8), where d_{ij} is the length of the shortest path from i to j and $n = n_1 + n_2$ is the total number of nodes. The best (smallest) score possible occurs when the node has a tie to every other node, in which case the total distance to all others is $n-1$. Thus, closeness centrality is usually normalized by dividing c_i into $n-1$.

$$\begin{aligned} c_i &= \sum_j^n d_{ij}\\ c_i^* &= \frac{n-1}{c_i}\,. \end{aligned} \tag{9}$$

In the bipartite case, the maximum number of nodes that a node can be distance 1 from is the number of nodes in the other class. For other nodes in its own class, the closest it can be is two links away. Thus, the theoretical bipartite minimum value of c_i (where $i \in V_1$) is $n_2 + 2(n_1-1)$ and the minimum for c_j (where $j \in V_2$) is $n_1 + 2(n_2-1)$. Therefore, closeness centrality can be normalized in bipartite graphs as follows:

$$\begin{aligned} c_i^* &= \frac{n_2 + 2(n_1-1)}{c_i}\,, \quad \text{for } i \in V_1\,,\\ c_j^* &= \frac{n_1 + 2(n_2-1)}{c_j}\,, \quad \text{for } j \in V_2\,. \end{aligned} \tag{10}$$

Betweenness centrality refers to the sum of shares of shortest paths that pass through a given node. The betweenness of node k in an ordinary graph is defined by Eq. (11), where g_{ij} is the number of geodesic paths from node i to node j, and g_{ikj} is the number of geodesic paths from i to j that pass through k.

$$b_k = \frac{1}{2}\sum_{i\neq k}^n \sum_{j\neq k,i}^n \frac{g_{ikj}}{g_{ij}}\,. \tag{11}$$

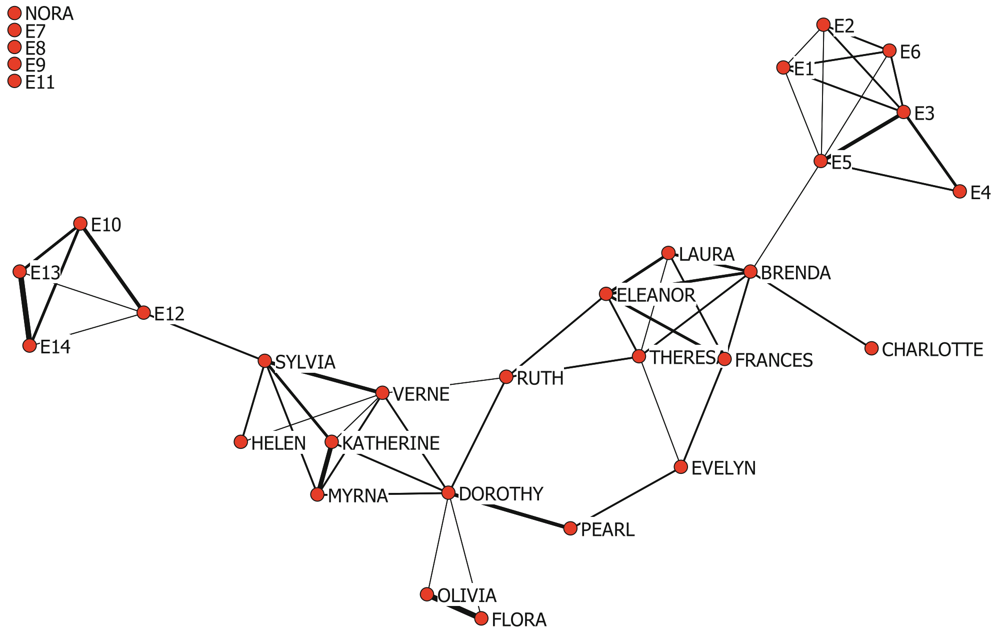
Betweenness is ordinarily normalized by dividing by $(n-1)(n-2) = n^2 - 3n + 2$, which is the maximum betweenness that any node can achieve in a graph with n nodes, which occurs for the node at the center of a star-shaped graph. This maximum is appropriate for bipartite graphs only when one mode has just one node; otherwise we must take account of the sizes of each vertex set. Eq. (12) gives the maximums for nodes in each vertex set as a function of the vertex set sizes. In the equation, x div y refers to integer division of x by y and x mod y refers to the remainder of an integer division of x by y.

$$\begin{aligned} b_{V_1\,\max} &= \tfrac{1}{2}\big[n_2^2(s+1)^2 + n_2(s+1)(2t-s-1)\\ &\qquad - t(2s-t+3)\big]\\ s &= (n_1-1) \operatorname{div} n_2\,, \quad t = (n_1-1) \mod n_2\\ b_{V_2\,\max} &= \tfrac{1}{2}\big[n_1^2(p+1)^2 + n_1(p+1)(2r-p-1)\\ &\qquad - r(2p-r+3)\big]\\ p &= (n_2-1) \operatorname{div} n_1\,, \quad r = (n_2-1) \mod n_1 \end{aligned} \tag{12}$$

Given these maxima, we can normalize standard betweenness centrality for bipartite graphs by dividing by these maxima, as shown in Eq. (13).

$$\begin{aligned} b_i^* &= \frac{b_i}{b_{V_1\,\max}}\,, \quad \text{for } i \in V_1\,,\\ b_j^* &= \frac{b_j}{b_{V_2\,\max}}\,, \quad \text{for } j \in V_2\,. \end{aligned} \tag{13}$$

Cohesive Subgroups Detecting cohesive subgroups is somewhat more difficult in bipartite graphs than in ordinary graphs. For example, one of the earliest formal definitions of a subgroup is the clique [11] which is defined



Social Network Analysis, Two-Mode Concepts in, Figure 11
Graph induced by Pearson correlations across biclique membership profiles. Edges indicate a correlation greater than 0.4

as a maximal complete subgraph – i. e., a set of nodes that is as large as possible subject to the condition that every pair is adjacent. However, in bipartite graphs, cliques are impossible, since in any connected triple, two nodes must be non-adjacent. Thus, cliques are not useful in this context. In addition, relaxations of the clique concept based on density or frequency of paths (such as k-plexes, lambda sets and ls-sets) do not work well, due to the sparse nature of bipartite graphs. For example, if we calculate k-plexes for the DGG dataset, we find huge numbers of very small groups. For $k = 2$, there are 394 2-plexes of size 3 or greater. For $k = 3$, there are 5,553 3-plexes, and for $k = 4$, there are 37,633 4-plexes of size 3 or greater.

However, other classical notions of cohesive subgroups make more sense for bipartite graphs. In particular, relaxations of the clique concept based on distance, such as n-cliques, n-clans and n-clubs have good interpretations in bipartite graphs and work well in practice. In fact, 2-cliques defined on the bipartite graph have the property of bipartite completeness, meaning that, within the bipartite subgraph induced by the 2-clique, every node of one mode has a tie with every node of the other mode, which is to say that all possible ties are present. In this sense, for bipartite graphs, 2-cliques capture the underlying idea of a clique better than cliques do. This notion has been formalized in the definition of a *biclique*, which is defined as a maximal complete bipartite subgraph. Mathematically, it is identical to a 2-clique computed on the bipartite representation.

As a practical example, a biclique analysis of the DGG dataset finds 68 bicliques, and a secondary analysis of similarities of nodes across biclique membership profiles does a good job of revealing structure, as shown in Fig. 11. In the figure, an edge is shown between two nodes if the Pearson correlation between their biclique membership profiles is greater than 0.4 (i. e., $(u, v) \in E$ iff $r_{ij} > 0.4$). The two groups are women are clearly shown, as well as the two groups of events that are associated with each of the groups of women. In addition, the separation of Flora and Olivia is clearly shown, as well as the bridging position of Ruth and Pearl.

The success of bicliques as 2-mode analogues of cliques suggests the possibility of creating 2-mode analogues for other cohesive subgroup concepts that we previously dismissed as unusable in the 2-mode context. For example, a *k-plex* is defined in ordinary network analysis as a maximal subgraph $S(V, E)$ of size n such that each member of the k-plex is adjacent to at least $n - k$ others. By analogy, for 2-mode networks we can define a (k_1, k_2)-*biplex* as a maximal bipartite subgraph $G(V_1, V_2, E)$ such that each node in V_1 is adjacent to $|V_2| - k_2$ others and each node in V_2 is adjacent is adjacent to $|V_1| - k_1$ others. Obviously, a (0,0)-biplex is a biclique. A useful feature of this relaxation of a biclique is that we can set different standards of cohesiveness for each mode of the data network, perhaps reflecting the relative sizes of the modes, or the affiliative capabilities of the nodes themselves.

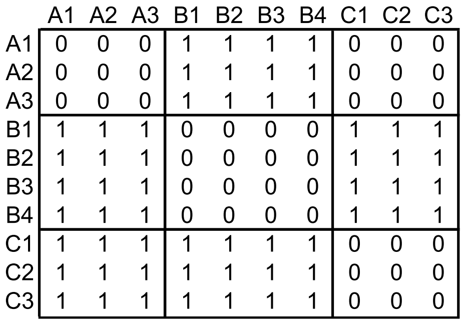
Bimodal Approaches to Positions and Roles Two concepts of position and role are particularly well known in social network analysis. These are structural equivalence and regular equivalence. Generalizations to 2-mode (and higher) data were developed by Borgatti [3] and Borgatti and Everett [4] and Everett and Borgatti [7]. We consider each of these in turn.

Structural Equivalence In the best definition of structural equivalence [8], two nodes u and v are said to be structurally equivalent if there exists a graph automorphism that would be an identity except that u and v are mapped to each other. In other words, given a diagram of the graph, you could swap nodes u and v (and no other nodes) without changing the structure of the network one iota. A key implication of this definition is that structurally equivalent nodes have identical relational environments – aside from each other, they are connected and not connected to exactly the same third parties. As a result, one approach to identifying structurally equivalent nodes is to compute a similarity measure among rows and columns of the adjacency matrix defining the graph.

This approach requires no modification for the bipartite case. In fact, the unimodal analysis described in Sect. "Unimodal Approaches to Two-Mode Data" (applied to each mode in turn) is precisely the same as computing structural equivalence on the 2-mode incidence matrix (Fig. 1). In addition, for certain measures such as the Jaccard coefficient, computing structural equivalence on the bipartite adjacency matrix B of Fig. 8 gives exactly the same results as computing it on the 2-mode incidence matrix X.

Another approach to structural equivalence is known as blockmodeling. Instead of (or as a result of) measuring the extent of structural equivalence between nodes, we partition the nodes into classes such that nodes in the same class are structurally equivalent (or nearly so). Given such a partition of nodes, we can reorder and partition the corresponding rows and columns of the adjacency matrix. This in turn induces to partition of matrix cells into *matrix blocks*. A characteristic property of perfect structural equivalence partitions is that matrix blocks are necessarily homogeneous with respect to cell values, and each block will consist of all 1s or all 0s (known as 1-blocks and 0-blocks respectively). Figure 12 gives an example with three equivalence classes.

In the 2-mode case, one approach is to blockmodel the bipartite adjacency matrix B. This can be done, but the bipartite structure imposes certain constraints. For example, blocks involving within-mode ties must be 0-blocks. In addition, the best 2-block partition will almost certainly be the mode partition (except in trivial cases), and in general, all other partitions will be refinements of the mode partition (i. e., they will be nested hierarchically within the mode partition).



	A1	A2	A3	B1	B2	B3	B4	C1	C2	C3
A1	0	0	0	1	1	1	1	0	0	0
A2	0	0	0	1	1	1	1	0	0	0
A3	0	0	0	1	1	1	1	0	0	0
B1	1	1	1	0	0	0	0	1	1	1
B2	1	1	1	0	0	0	0	1	1	1
B3	1	1	1	0	0	0	0	1	1	1
B4	1	1	1	0	0	0	0	1	1	1
C1	1	1	1	1	1	1	1	0	0	0
C2	1	1	1	1	1	1	1	0	0	0
C3	1	1	1	1	1	1	1	0	0	0

Social Network Analysis, Two-Mode Concepts in, Figure 12
Perfect structural equivalence blockmodel

A more elegant (and computationally efficient) approach is to work from the 2-mode incidence matrix X. In this case, we redefine a blockmodel to refer to a coordinated pair of partitions, one for the rows and one for the columns. We then redefine structural equivalence as follows. Given a 2-mode matrix X with modes V_1 and V_2, a 2-mode blockmodeling consists of a pair of partitions P_1 and P_2 such that (a) for all u, v in V_1, $p_1(u) = p_1(v)$ iff $x_{\text{uw}} = x_{\text{vw}}$ for all w in V_2 and (b) for all a, b in V_2, $p_2(a) = p_2(b)$ iff $x_{\text{za}} = x_{\text{zb}}$ for all z in V_1. (The notation $p(u)$ indicates the equivalence class that node u belongs to in partition P.) Restating this in words, a 2-mode structural equivalence blockmodeling is one in which row nodes in the same class if and only if they have identical rows, and column nodes are in the same class if and only if they have identical columns. An example involving 4 classes of rows and 3 classes of columns is shown in Fig. 13.

In empirical work, of course, we do not expect to obtain perfect 1-blocks and 0-blocks. Instead we seek partitions that minimize the number of errors (where we define an error as either a 1 inside a 0-block or a 0 inside a 1-block).

Regular Equivalence Two nodes are said to be *regularly equivalent* (i. e., play the same structural role) to each other to the extent that they have ties (lack of ties) to corresponding others who are themselves regularly equivalent to each other. That is, u and v are regularly equivalent

	E1	E2	E3	F1	F2	F3	F4	G1	G2	G3
A1	1	1	1	1	1	1	1	0	0	0
A2	1	1	1	1	1	1	1	0	0	0
A3	1	1	1	1	1	1	1	0	0	0
B1	1	1	1	0	0	0	0	0	0	0
B2	1	1	1	0	0	0	0	0	0	0
B3	1	1	1	0	0	0	0	0	0	0
B4	1	1	1	0	0	0	0	0	0	0
C1	0	0	0	1	1	1	1	0	0	0
C2	0	0	0	1	1	1	1	0	0	0
C3	0	0	0	1	1	1	1	0	0	0
D1	0	0	0	1	1	1	1	1	1	1
D2	0	0	0	1	1	1	1	1	1	1

Social Network Analysis, Two-Mode Concepts in, Figure 13
A 2-mode structural equivalence blockmodel in which the row partition has 4 classes and the column partition has 3 classes

if, (a) for all x such that u has a tie to x, there exists a y that is regularly equivalent to x such that v has a tie to y, and (b) for all x such that u has an incoming tie from x, there exists a y that is regularly equivalent to x such that v has an incoming tie from y. Whereas in structural equivalence equivalent nodes are connected to the same others, in regular equivalence equivalent nodes are connected to the same types of others.

Alternatively, we can define regular equivalence in terms of a partition C of nodes into labeled classes (known as colors) such that regularly equivalent nodes are required to have the same colors in their neighborhoods. That is, for all nodes u and v,

$$C(u) = C(v) \text{ implies } C(N(u)) = C(N(v)) , \qquad (14)$$

where $C(N(u))$ is the set of distinct colors found among the nodes constituting u's immediate neighborhood.[1] Taking a blockmodeling perspective, it can readily be seen that Eq. (14) implies a blockmodel in which each matrix block is either entirely zero (a "zeroblock") or contains at least one 1 in every column and in every row (a "regular oneblock"), as illustrated in Fig. 14.

Equation (14) lends itself easily to the bipartite case. Instead of seeking a single partition of nodes, we seek a different partition for each mode. Labeling these partitions R and C, we modify our definition such that for all $u, v \in V_1$ and $x, y \in V_2$,

$$\begin{aligned} C(u) = C(v) \text{ implies } R(N(u)) = R(N(v)) , \quad \text{and} \\ R(x) = R(y) \text{ implies } C(N(x)) = C(N(y)) . \end{aligned} \qquad (15)$$

	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10
N1	0	0	0	1	0	0	1	0	0	0
N2	0	0	0	0	1	0	0	0	0	0
N3	0	0	0	1	1	1	1	0	0	0
N4	1	0	0	0	0	0	0	0	1	1
N5	1	0	1	0	0	0	0	1	1	0
N6	0	1	0	0	0	0	0	0	1	1
N7	0	1	1	0	0	0	0	1	0	1
N8	1	0	1	0	1	1	0	0	0	0
N9	0	0	1	0	0	1	0	0	0	0
N10	0	1	0	1	1	0	1	0	0	0

Social Network Analysis, Two-Mode Concepts in, Figure 14
Perfect regular equivalence blockmodel on an ordinary graph

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
C1	1	0	1	0	1	1	0	0	0	0
C2	0	0	1	0	0	1	0	0	0	0
C3	0	1	0	1	1	0	1	0	0	0
C4	1	0	0	0	0	0	0	0	1	1
C5	1	0	1	0	0	0	0	1	1	0
C6	0	1	0	0	0	0	0	0	1	1
C7	0	1	1	0	0	0	0	1	0	1
C8	0	0	0	0	1	1	0	0	0	0
C9	0	0	0	0	0	1	0	0	0	0
C10	0	0	0	1	1	0	1	0	0	0
C11	1	0	1	0	0	0	0	0	1	1
C12	0	1	0	0	0	0	0	1	0	1

Social Network Analysis, Two-Mode Concepts in, Figure 15
A 2-mode regular equivalence blockmodel

Of course, the neighbors N() of any node consist entirely of nodes belonging to the other mode. Thinking in terms of the 2-mode incidence matrix shown in Fig. 1, Eq. (15) implies that we can section the matrix into rectangular blocks such that each block is zeroblock or a regular 1-block. An example of a 2-mode regular blockmodel is shown in Fig. 15.

As with structural equivalence, in empirical work we do not expect to obtain perfect 1-blocks and 0-blocks. Instead we seek partitions induce matrix blocks that are as nearly regular as possible.

Conclusion

In this chapter we have considered methods of visualizing and analyzing 2-mode network data. An examination of the chapter suggests that four strategies are employed for handling 2-mode data. First, there is converting the data into two separate 1-mode datasets, and analyzing each separately. Second, there is using existing 1-mode methods on a bipartite representation of the 2-mode data, and essentially ignoring its special features. Third, there is using 1-mode methods but either adjusting the interpretation or

[1] For directed graphs, we require $C(N^o(u)) = C(N^o(v))$ and $C(N^i(u)) = C(N^i(v))$, where $N^o(v)$ refers to the set of nodes that v sends a tie to, and $N^i(u)$ refers to the set of nodes that v receives a tie from.

applying some kind of normalization to adjust the results. Finally, there is developing new methods designed specifically for the 2-mode case. The chapter has provided examples of all four strategies.

It is worth reminding the reader that only 2-mode data viewed in a network context has been considered here. In principle, most data dealt with by social scientists is structured as a 2-mode matrix of cases (rows) and variables (columns) and most statistical techniques assume such data. This chapter obviously does not discuss this vast pantheon of techniques, although it is clear that many would be appropriate, particularly structure-finding techniques such as factor analysis and correspondence analysis.

Future Directions

Although it would be a mistake to think of 2-mode data as an advance over 1-mode data, it is important to note that there are many cases were extending network analysis methodology to more than 2 modes is desirable. For example, we might analyze membership in organizations over time, yielding a 3-way, 3-mode data matrix of relations that is person by organization by period. Similarly, we might be interested in modeling the intellectual landscape of an academic field, representing publications as k-mode bundles of authors, journals, years, institutional affiliations and topics.

Bibliography

1. Bonacich P (1972) Techniques for analyzing overlapping memberships. In: Costner HL (ed) Sociological methodology. Jossey-Bass, San Francisco, pp 176–185
2. Bonacich P (1991) Simultaneous group and individual centralities. Soc Netw 13:155–168
3. Borgatti SP (1989) Regular equivalence in graphs, hypergraphs, and matrices. Doctoral Dissertation, Univ of California, Irvine, Ann Arbor
4. Borgatti SP, Everett M G (1992) Regular blockmodels of multiway, multimode matrices. Soc Netw 14:91–120
5. Borgatti SP, Everett MG (1997) Network analysis of 2-mode data. Soc Netw 19(3):243–269
6. Davis A, Gardner B, Gardner M (1941) Deep South: A Social Anthropological Study of Caste and Class. University of Chicago Press, Chicago
7. Everett MG, Borgatti SP (1993) An extension of regular colouring of graphs to digraphs, networks and hypergraphs. Soc Netw 15:237–254
8. Everett MG, Borgatti SP (1994) Regular equivalence: General theory. J Math Sociol 19(1):29–52
9. Faust K (1997) Centrality in affiliation networks. Soc Netw 19:157–191
10. Kamada T, Kawai S (1989) An algorithm for drawing general undirected graphs. Inf Process Lett 31:7–15
11. Luce RD, Perry AD (1949) A method of matrix analysis of group structure. Psychometrika 20:319–327

Social Networks, Algebraic Models for

PHILIPPA PATTISON
Department of Psychology, University of Melbourne, Parkville, Australia

Article Outline

Glossary

Social network A *social network* comprises a set of relationships among members of a set $N = \{1, 2, \dots, n\}$ of actors. It can be represented by an $n \times n$ binary array X recording the presence or absence of a social relationship, or *tie*, between each pair (or ordered pair) of members of $N = \{1, 2, \dots, n\}$. If there is a relationship from actor k to actor l, we write $X(k, l) = 1$; otherwise, $X(k, l) = 0$. If the relationship is a property of a pair of actors, the network is *nondirected*; if it is a property of an ordered pair, the network is termed *directed*. The directed network X may also be regarded as a *binary relation* R_X on the set N with $(k, l) \in R_X$ if and only if $X(k, l) = 1$; equivalently, it may be construed as a *directed graph* with *node set* N and *arc set* R_X, with an *arc* from node k to node l if and only if $(k, l) \in R_X$.

Affiliation network An *affiliation network* is an $n \times g$ binary array X recording the membership of each of a set N of actors in a prescribed set G of *groups*, with $X(k, l) = 1$ if actor k is a member of group l, and $X(k, l) = 0$ otherwise.

Multiple network A *multiple network* is a collection of networks for each of a set of r relations. We let $X_m(k, l) = 1$ if the tie from k to l corresponding to the relation of *type m* is present; and $X_m(k , l) = 0$ if the tie is absent. Nodes k and l are joined by a *labeled walk* with *label* $Y_1 Y_2 \dots Y_j$ if there is a sequence of nodes $k = k_0, k_1, \dots, k_j = l$, for which $Y_h(k_{h-1}, k_h) = 1$ for $h = 1, 2, \dots, j$.

Local network The *(1-)neighborhood* of a subset P of actors in a network is defined to be the set $P \cup \{l \in N\colon X_m(k,l) = 1$ for some $k \in P$ and some relation $m\}$. The *q-neighborhood* of P is then defined recursively as the 1-neighborhood of the $(q-1)$-neighborhood of P. The *q-local network* of the subset P of N is the network restricted to its q-neighborhood.

Algebra A (*partially ordered*) *algebra* is a triple $[S, F, \leq]$, where S is a nonempty set of elements (usually assumed to be finite), F is a specified set of operations, f_α, each mapping a power $S^{n(\alpha)}$ of S into S, for some non-negative finite integer $n(\alpha)$, and $\leq$ is a partial order on S. Each operation f_α is assumed to be *isotone* in each of its variables: that is, if $x_i \leq y_i$ $(x_i, y_i \in S;\ i = 1, 2, \ldots, n(\alpha))$, then $f_\alpha(x_1, x_2, \ldots, x_{n(\alpha)}) \leq f_\alpha(y_1, y_2, \ldots, y_{n(\alpha)})$. A *family of algebras* is a collection of algebras each having the same set F of operations and satisfying a specified set of postulates. Two algebras belonging to the same family are termed *similar*.

Partial algebra A *partial algebra* is a triple $[S, F, \leq]$, where S is a nonempty set of elements (usually assumed to be finite), F is a specified set of *partial* operations, f_α, each mapping some subset $T^{(\alpha)}$ of $S^{n(\alpha)}$ into S, for some non-negative finite integer $n(\alpha)$, and $\leq$ is a partial order on S. Each partial operation f_α is assumed to be *isotone* in each of its variables: that is, if $x_i \leq y_i$ $(x_i, y_i \in S;\ i = 1, 2, \ldots, n(\alpha))$, then $f_\alpha(x_1, x_2, \ldots, x_{n(\alpha)}) \leq f_\alpha(y_1, y_2, \ldots, y_{n(\alpha)})$ provided that both $(x_1, x_2, \ldots, x_{n(\alpha)})$, $(y_1, y_2, \ldots, y_{n(\alpha)}) \in T^{(\alpha)}$. A *family of partial algebras* is a collection of algebras each having the same set F of partial operations defined on the same subsets $T^{(\alpha)}$ of the power sets $S^{n(\alpha)}$.

Semigroup A *(partially ordered) semigroup* is an algebra $[S, F, \leq]$ in which F comprises a single binary operation f satisfying the *associativity* condition:

$$f(f(x,y),z) = f(x, f(y,z))\,.$$

Lattice A *lattice* L is an algebra $[S, F, \leq]$ in which F comprises two associative and commutative binary operations, $\wedge$ and $\vee$ (termed *meet* and *join*, respectively) satisfying the identities:

$$x \wedge x = x\,, \quad x \vee x = x$$

and

$$x \wedge (x \vee y) = x \vee (x \wedge y) = x\,.$$

The operations are isotone, so that $x \leq y$ is equivalent to the pair of conditions:

$$x \wedge y = x \quad \text{and} \quad x \vee y = y\,,$$

and the operations of meet and join may be interpreted as the greatest lower bound and least upper bound, respectively. A lattice L is *distributive* if the identity

$$x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z)$$

holds. A lattice L is *modular* if, whenever $x \leq z$, then

$$x \vee (y \wedge z) = (x \vee y) \wedge z\,.$$

Role algebra A *role algebra* is an algebra $[S, F, \leq]$ in which F comprises a single binary composition operation satisfying the condition: $s \leq t$ in S implies $su \leq tu$ in S, for any $u \in W$.

Homomorphism An (*isotone*) *homomorphism* from an algebra $A = [S, F, \leq]$ onto a similar algebra $B = [T, F, \leq]$ is a mapping $\phi\colon S \to T$ such that,

(i) for all $f_\alpha \in F$ and $x_i \in S$,

$$\phi(f_\alpha(x_1, x_2, \ldots, x_{n(\alpha)})) = f_\alpha(\phi(x_1), \phi(x_2), \ldots, \phi(x_{n(\alpha)}))\,; \quad \text{and}$$

(ii) $x \leq y$ in S implies $\phi(x) \leq \phi(y)$ in T.

The algebra B is termed a (*homomorphic*) *image* of A, and we write $B = \phi(A)$. Each homomorphism ϕ from $A = [S, F, \leq]$ onto $B = [T, F, \leq]$ has a corresponding binary relation π on S (termed here a *π-relation*) in which $(x, y) \in \pi$ if and only if $\phi(y) \leq \phi(x)$. The equivalence relation σ_π defined by $(x, y) \in \sigma_\pi$ if and only if $(x, y) \in \pi$ and $(y, x) \in \pi$ is termed a *congruence relation*.

Homomorphism lattice The *homomorphism lattice* $L(A)$ of the algebra $A = [S, F, \leq]$ is the collection of all homomorphisms of A partially ordered by the relation: $\phi_1 \leq \phi_2$ if, for all $x, y \in S$, $\phi_2(x) \leq \phi_2(y)$ implies $\phi_1(x) \leq \phi_1(y)$. The *lattice $L_\pi(A)$ of π-relations* on A, dual to $L(A)$, has the partial ordering: $\pi_1 \leq \pi_2$ if $(x, y) \in \pi_1$ implies $(x, y) \in \pi_2$, for any $x, y \in S$.

Definition of the Subject

Algebraic models have been proposed to represent structure in social networks. They are usually constructed from a set of operations and relations defined on network constituents such as paths or walks in the network, or vectors of ties directed to or from individual network members. The algebra represents the relationships among these constituents, for example, relations of ordering among all possible walks in a multiple network, or relations of overlap and ordering among profiles of group membership. To date, they have been used to represent kinship structures [6,9,48], role structures and stability in multiple networks [1,5,8,36], states of diffusion processes in networks [30], informal hierarchy [17], connectivity structures [13] and the structure of membership in affiliation networks [16,38]. In several cases, partial algebraic representations have also been proposed [35,39,40].

Algebraic models have usually been proposed as exact structural representations and, in many cases, therefore, their application to network data has been coupled with some form of prior data aggregation such as blockmodeling [5,50]; see also the entry on ▶ Positional Analysis and Blockmodeling. More rarely, algebraic construction and data aggregation have been combined into a single step [8,33].

Introduction

Algebraic models for relational structures such as networks have their foundations in anthropology and were developed for general network structures by Lorrain, White and colleagues [5,26,27]. These general network representations based on path structures in networks emerged from more specific models for path structures in kinship systems, including permutation group models for Australian aboriginal kinship systems [48]. Boyd extended this approach in an important way when he proposed that the structural evolution of kinship systems could be analyzed in terms of homomorphic mappings [6]; applications to network semigroups were described later [5,37].

Lattice representations for relational structures have a long history in mathematics [2] and applications to relational data in many domains have been developed extensively by Wille and collaborators in formal concept analysis [19,20,51]. They have also been applied to affiliation network structures [16,45,47].

The form of algebraic structure regarded as appropriate in a particular context is a matter for substantive consideration. The large majority of forms that have been proposed to date have depended upon the importance of paths or walks in the network, the symmetric or asymmetric nature of network ties, and the patterns of overlap among tie partners. A focus on one or more of these features has resulted in algebraic groups, semigroups, partially ordered semigroups, local role algebras and lattices as representations of different aspects of structure in multiple social networks and affiliation networks [5,8,14,16,29,36,48] and it is on these forms that we focus here.

Algebras from Networks

A Language for Discriminating Among Relations

We begin with a finite set $\Sigma = \{x_1, x_2, \ldots, x_r\}$ of r relational terms. We term Σ a relational *alphabet* and call each term x_i in Σ a *letter* of the alphabet. The letters $x_1, x_2, \ldots$ denote relational terms such as "friendship", "co-worker", "confidant", and so on. If $r = 1$, the set Σ simply identifies a single network relation.

In order to construct algebraic representations that are sensitive to social structural form, claims in the social network literature can be used to introduce a *free word algebra* on the alphabet Σ. The algebra is intended to provide a language for describing more complex relational terms that can be constructed from the primitive set Σ and so to provide the means for discriminating among social structural forms. The claims about appropriate ways of constructing new terms from the primitive set may be drawn from a number of sources, but one set of such claims is helpfully laid out by White [49]. White argues that there are three universal ways in which types of tie (i. e., relational terms) are discriminated. The first is in terms of patterns of overlap; the second in terms of asymmetry of relation; and the third in terms of the institutionalization of indirect ties. The overlap of two relational ties refers to their joint occurrence and can arguably be described using an *intersection* operation, $\cap$. The asymmetry of a tie can be described by the relationship between the tie and its *converse*, that is the coding of whether, for two individuals k and l, there is a tie from k to l and from l to k. The tie is *symmetric* if the tie from k to l occurs only in conjunction with the tie from l to k, and the tie is *strictly asymmetric* if a tie from k to l occurs only in the absence of a tie from l to k. We use the expression x' to denote the converse of the relational term x. Indirect ties may be described by a *composition*, or *concatenation* operation, denoted $^\circ$. If x_1 refers to the term "friendship" and x_2 to "coworker", then the concatenation $x_1{}^\circ x_2$ may be used to refer to a relation that holds between a person and his or her friend's co-workers.

We denote by F the set of operations from which our descriptive language is to be built. Unless otherwise specified, we assume here that F comprises the three operations just described; that is $F = \{\cap, {}', {}^\circ\}$. Each operation f_α in F possesses an *arity*, $n(\alpha)$, that determines the form of new terms created from Σ by f_α: in particular, f_α leads to expressions of the form $f_\alpha(y_1, y_2, \ldots, y_{n(\alpha)})$; $y_m \in \Sigma$, $m = 1, 2, \ldots, n(\alpha)$. The operations $\cap$ and $^\circ$ have arity 2, that is, they are *binary* and lead to expressions that can be written as $y_1 \cap y_2$, and $y_1{}^\circ y_2$, respectively, whereas the converse operation $'$ is *unary* (that is, it has an arity of 1) and leads to an expression that can be denoted y_1' (where $y_1, y_2 \in \Sigma$).

We describe a recursive construction of a free word algebra $W = [\Sigma, F]$ from the alphabet Σ and the set F of operations [2]. We call each letter in Σ an F-polynomial of *rank* 0. For any positive integer ρ, we define an F-polynomial of rank ρ recursively as an expression, termed a *word*, of the form $f_\alpha(u_1, \ldots, u_{n(\alpha)})$, where each u_j is an F-polynomial of rank $\leq \rho - 1$ (and at least one u_j is an F-polynomial of rank $\rho - 1$). The *(free) word algebra* $W = [\Sigma, F]$ comprises all distinct F-polynomials of finite rank (where $f_\alpha(u_1, \ldots, u_{n(\alpha)})$ and $f_\beta(v_1, \ldots, v_{n(\beta)})$ are distinct unless $\alpha = \beta$ and $u_k = v_k$ for all $k = 1, \ldots, n(\alpha)$).

It is sometimes convenient to restrict attention to the collection of words of rank no greater than some fixed integer ρ. So we also define $W_\rho = [\Sigma, F]_\rho$ as the subset of the free word algebra $W = [\Sigma, F]$ comprising words of rank $\leq \rho$.

For example, suppose that F comprises the operations $\{\cap, ', ^\circ\}$ and that $\Sigma = \{x, y\}$. Then the elements of rank 0 in the word algebra $W = [\Sigma, F]$ are $\{x, y\}$; the elements of rank 1 are $\{x \cap x, x \cap y, y \cap x, y \cap y, x', y', x^\circ x, x^\circ y, y^\circ x, y^\circ y\}$; elements of rank 3 include $x \cap x \cap x, x \cap (x^\circ x), (x^\circ x) \cap x, x \cap x', x' \cap x,$ etc; and so on. In this case, the free word algebra $W = [\Sigma, F]$ comprises all composite relational terms that can be constructed from the primitive terms x and y using the intersection, converse and concatenation operations.

The free word algebra simply provides a language that permits certain forms of discrimination among types of tie to be encoded. It makes no structural assumptions whatever, although we might anticipate from the intended interpretation of the terms in F that certain terms in W will not need to be discriminated. For instance, we might suppose that: the intersection operation can be assumed to be *idempotent* ($x \cap x = x$, for all $x \in W$), *commutative* ($x \cap y = y \cap x$, for all $x, y \in W$) and *associative* ($(x \cap y) \cap z = x \cap (y \cap z)$, for all $x, y, z \in W$); that the converse operation satisfies $x'' = x$, for all $x \in W$; and that the concatenation operation is also associative. Rather than impose such postulates at this abstract level, though, we move directly to the evaluation of the terms in W for an empirical realization of a (multiple) network.

Algebras for Multiple Networks

The language provided by the free algebra $W = [\Sigma, F]$ can be used to determine all possible discriminations among ties that our structural claims suggest as appropriate. To describe the structural forms to which observed network ties give rise, each of the derived relations for an observed multiple network can be computed. There is a straightforward correspondence between the words u in $W = [\Sigma, F]$ expressed in terms of letters in Σ and operations in F, on the one hand, and Boolean matrix manipulations of the binary constituents in the multiple network, on the other. The correspondence is achieved by replacing:

- the relational term x_m by the array X_m;
- the intersection operation $\cap$ by Boolean intersection of arrays (i. e., element-wise Boolean multiplication, denoted $\bullet$); and
- the composition operation $^\circ$ by Boolean matrix multiplication.

Boolean intersection, transposition and composition are defined for $n \times n$ binary arrays Y_1 and Y_2 by:

- $Y_1 \bullet Y_2(k, l) = Y_1(k, l) Y_2(k, l)$;
- $Y_1'(k, l) = Y_1(l, k)$; and
- $Y_1 Y_2(k, l) = Y_1(k, 1) Y_2(1, l) + Y_1(k, 2) Y_2(2, l) + \cdots + Y_1(k, n) Y_2(n, l)$.

The term $u \in W$ then corresponds to an $n \times n$ Boolean array, denoted X_u, with $X_u(k, l) = 1$ if and only if there is a relational tie described by the word u from k to l. We also define a partial order on the set of relations $\{X_u : u \in W\}$ by:

$$X_u \leq X_v \quad \text{if and only if} \quad X_u(k, l) \leq X_v(k, l) \quad \text{for all } k,\ l \in N\ .$$

Suppose that we evaluate X_u for all $u \in W$. We define a generalized version of the *Axiom of Quality* by setting $u \leq v$ in W whenever $X_u \leq X_v$ [5].

Defining an induced partition θ on W by $(u, v) \in \theta$ if and only if $u \leq v$ and $v \leq u$ and letting $S = W/\theta$ be the set of associated classes with the induced partial order:

$$s \leq t \text{ in } S \quad \text{if } u \in s\ ,\ v \in t \text{ and } u \leq v \text{ in } W\ ,$$

we can then define a general (partially ordered) algebraic structure on S.

A (*partially ordered*) *algebra* is a triple $[S, F, \leq]$, where S is a nonempty set of elements (usually assumed to be finite), F is a specified set of operations, f_α, each mapping a power $S^{n(\alpha)}$ of S into S, for some non-negative finite integer $n(\alpha)$, and $\leq$ is a partial order on S. Each operation f_α is assumed to be *isotone* in each of its variables: that is, if $x_i \leq y_i$ ($x_i, y_i \in S$; $i = 1, 2, \ldots, n(\alpha)$), then $f_\alpha(x_1, x_2, \ldots, x_{n(\alpha)}) \leq f_\alpha(y_1, y_2, \ldots, y_{n(\alpha)})$. A *family of algebras* is a collection of algebras each having the same set F of operations and satisfying a specified set of postulates. Two algebras belonging to the same family are termed *similar*.

For $F = \{\circ\}$, the algebra $[S, F, \leq]$ is a *partially ordered semigroup* [36].

The Semigroup of a Multiple Network

For multiple networks, a composite relation constructed as the concatenation of observed relations records the presence of a particular form of *labelled walk*. Specifically, actor k is connected to actor l by a walk with label u if and only if $X_u(k, l) = 1$. If $X_u \leq X_v$, two actors who are joined by a walk with label u are also necessarily joined by a walk with label v.

An algorithm for constructing the semigroup of a multiple network $[S, \{\circ\}, \leq]$ is as follows:

(S1) Let $W_1 = \{X_1, X_2, \ldots, X_r\}$ and set $i = 1$.
(S2) Construct the networks with edge sets WZ for each $W \in W_i$ and each $Z \in W_1$. Place WZ in W_{i+1} if $WZ \neq Y$, for any $Y \in W_j$, $j = 1, 2, \ldots, i+1$; otherwise, $WZ = Y$ for some $Y \in W_j$ $(j = 1, 2, \ldots, i+1)$ is an *equation* in the semigroup S.
(S3) If $W_{i+1} = \emptyset$, stop and compute the partial order among elements in $W_1 \cup W_2 \cup \ldots$; otherwise, set i to i+1 and return to step S2.

The equations $WZ = Y$ may be recorded in a *right multiplication table* in which Y appears as the table entry corresponding to the row labelled W and the column labelled Z. The partial ordering on S is constructed from the collection of networks in $W_1, W_2, \ldots$.

Lattices from Networks

Lattice representations for networks arise in two important ways. In the first, elements of the lattices are elements of $S = W/\theta$ for $F \supseteq \{\cap, \circ\}$. The meet of two elements is their intersection, and their join is the least relation in S which is greater than or equal to both of them. Indeed, the algebra $[S, F, \leq]$ is both a lattice and a partially ordered semigroup in this case.

In the second type of lattice construction, the elements correspond to rows or columns of an affiliation network. Since the affiliation array provides an explicit representation of the co-constitution of groups by actors and actors by groups, the lattice structures to which it gives rise can be used to analyze the duality of actors and groups [10,15,16,31,33,38,47].

A lattice denoted $L(X)$ is generated by the rows of a matrix X under the intersection operation; it is termed the *Galois* or *concept lattice* of X [20]. In $L(X)$, the meet of two elements is equal to their intersection, while their join is the minimal element in the lattice which is greater than or equal to both of them. A lattice can also be constructed from the rows of the matrix X under the union operation. In this lattice, termed the *Zareckii lattice* by Boyd, the join of two vectors is equal to their union, and their meet is the maximal vector which is less than or equal to both of them [8].

To construct the Galois lattice, one simply needs to compute all distinct intersections among rows of X, add the vector $[1\ 1 \ldots\ 1]$ and compute the partial order among the resulting set of row vectors.

Role Algebras from Networks

Role algebras have been proposed as local network analogues of the semigroup representation of relational structure in an entire social network. The representation has been presented in several slightly different forms but each is derived from an original formulation proposed by Mandel [11,29,53,54].

Denote by $k * X_u$ the vector $[X_u(k, 1), X_u(k, 2), \ldots, X_u(k, n)]$ for $u \in W$. The vector indicates whether there is a relationship of type u from actor k to each other actor in the network. As before, we can envisage evaluating $k * X_u$ for all $u \in W$. We then define a k-centered version of the *Axiom of Quality* by setting

$$u \leq v \text{ in } W \quad \text{whenever} \quad k * X_u \leq k * X_v$$

and an induced partition θ on W by $(u, v) \in \pi$ if and only if $u \leq v$ and $v \leq u$. Also as before, we let $S = W/\theta$ denote the set of associated classes with the induced partial order:

$$s \leq t \text{ in } S \quad \text{if } u \in s\,,\ v \in t \text{ and } u \leq v \text{ in } W\,.$$

The algebra $[S, F, \leq]$ is termed a *role algebra*.

The local role algebra for actor k may be constructed as follows.

(RA1) Let $W_1 = \{X_1, X_2, \ldots, X_r\}$ and set $i = 1$.
(RA2) Construct the vectors $k * WZ$ for each $W \in W_i$ and each $Z \in W_1$. Place WZ in W_{i+1} if $k * WZ \neq k * Y$, for any $Y \in W_j$, $j = 1, 2, \ldots, i + 1$; otherwise, $k * WZ = k * Y$ for some $Y \in W_j$ $(j = 1, 2, \ldots, i + 1)$ and $WZ = Y$ is an *equation* in the role algebra.
(RA3) If $W_{i+1} = \emptyset$, stop and compute the partial order among elements in $W_1 \cup W_2 \cup \ldots$; otherwise, set i to i+1 and return to step RA2.

Partial Algebras

To achieve algebraic closure, it may be necessary to compute relations corresponding to words of high rank in W and, in the case of some representations at least, it can be argued that some restriction on these repeated operations should be imposed. Mandel, for instance, made such an argument for the composition operation in social networks; he claimed that short network paths are likely to be more salient to the members of the network than longer ones [29]. These considerations can be formalized with the introduction of partial algebraic structures, whose elements are subject to rank restrictions on some or all of the operations used to construct the algebra.

Let $W_\rho = [\Sigma, F]_\rho$ be the subset of the free word algebra $W = [\Sigma, F]$ comprising words of rank $\leq \rho$. Define a partial order on the elements of W_ρ by the generalized Axiom of Quality:

$$u \leq v \text{ in } W_\rho \quad \text{whenever} \quad X_u \leq X_v \, .$$

The relation $\leq$ is clearly both *reflexive* ($u \leq u$ for all u) and *transitive* ($u \leq v$ and $v \leq w$ implies $u \leq w$ for any u, v, w), that is, a *quasi-order*. Further, each operation f_α is *isotone* in each of its variables whenever the result of the operation is contained in W_ρ; that is, if $x_i \leq y_i$ ($x_i, y_i \in W_\rho$; $i = 1, 2, \ldots, n(\alpha)$), then $f_\alpha(x_1, x_2, \ldots, x_{n(\alpha)}) \leq f_\alpha(y_1, y_2, \ldots, y_{n(\alpha)})$, provided that both $f_\alpha(x_1, x_2, \ldots, x_{n(\alpha)}), f_\alpha(y_1, y_2, \ldots, y_{n(\alpha)}) \in W_\rho$. Thus it follows that the partial ordering $\leq$ on W_ρ gives rise to a *partial* algebra in the following sense.

A *partial algebra* is a triple $[S, F, \leq]$, where S is a nonempty set of elements (usually assumed to be finite), F is a specified set of *partial* operations, f_α, each mapping some subset $T^{(\alpha)}$ of $S^{n(\alpha)}$ into S, for some non-negative finite integer $n(\alpha)$, and $\leq$ is a partial order on S. Each partial operation f_α is assumed to be *isotone* in each of its variables: that is, if $x_i \leq y_i$ ($x_i, y_i \in S$; $i = 1, 2, \ldots, n(\alpha)$), then $f_\alpha(x_1, x_2, \ldots, x_{n(\alpha)}) \leq f_\alpha(y_1, y_2, \ldots, y_{n(\alpha)})$ provided that both $(x_1, x_2, \ldots, x_{n(\alpha)}), (y_1, y_2, \ldots, y_{n(\alpha)}) \in T^{(\alpha)}$. A *family of partial algebras* is a collection of algebras each having the same set F of partial operations defined on the same subsets $T^{(\alpha)}$ of the power sets $S^{n(\alpha)}$.

Partial semigroup algebras may be constructed from labelled walks in a multiple network of some maximum length ρ [40]. For role algebras, the partial role algebra for a node k derived from W_ρ is the role algebra associated with the ρ-local neighborhood of k [39].

Algebraic Structure

In some cases, it is possible to analyze the structural properties of algebraic representations of networks.

The Semigroup of a (Single) Network Under Composition

We first consider the semigroup for a single network array X. We say that there is a *walk* from actor k to actor l if $X(k, l) = 1$. There is a walk of *length* p from k to l if there is a sequence of actors $k = k_0, k_1, \ldots, k_p = l$ such that $X(k_{i-1}, k_i) = 1$, for $i = 1, 2, \ldots, p$. It is readily seen that $X^p(k, l) = 1$ if and only if there is a walk of length p from k to l. The sequence $X,\ X^2, \ldots$ generated by the composition operation is therefore a sequence recording the presence of walks of different lengths between all pairs of network actors. The semigroup structure of this sequence therefore describes relations of equality and ordering among walks in the network.

The sequence can be described with the help of the following definition. Let p be the least integer such that

$$X^p = X^q \quad \text{for some} \quad q > p \, .$$

Also let $q = p + d$ ($d \geq 1$) be the least integer satisfying this relation. The integers p and d are termed the *index* and *period* of the network, respectively. It can readily be seen that the sequence has the form

$$X, \ldots, X^{p-1}, X^p, \ldots, X^{p+d-1}, X^p, \ldots, X^{p+d-1}, \ldots \, .$$

The set $\{X, X^2, \ldots, X^{p+d-1}\}$ defines a semigroup S of which $G = \{X^p, \ldots, X^{p+d-1}\}$ is a group (defined below).

It is natural to ask what is the relationship between properties of a network and its index and period.

For a strongly connected network X (that is, a network in which there is a walk of some length from each actor to each other actor), the period d is the greatest common denominator of all integers m such that m is the length of a cycle in X (that is, a walk with the same initial and final node). If X has more than one strong component, its period is the least common multiple of the periods of those strong components [24].

A network X with index 1 and period 1 is termed *idempotent*; for a network of index p and period 1, the network X^p is idempotent. It may be shown that every idempotent network is a pseudo-order, defined as follows [43]. Let Z be a quasi-order on N and let $e_Z = Z \cap Z'$. An actor $k \in N$ is termed *Z-strict* if there is no distinct actor $l \neq k$ such that $Z(k, l) = 1$. The subset H of N is *Z-permissible* if each of its actors is Z-strict and there are no pairs $k, l \in N$ such that k covers l or l covers k (recall that k covers l if $Z(k, j) = 1$ and $Z(j, l) = 1$ implies $j = k$ or $j = l$). Define the relation Z_H by

$$Z_H(k, l) = 1 \quad \text{iff} \quad k = l \text{ and } k \in H \, .$$

Then a network X is a *pseudo-order* if

$$X = Z \backslash Z_H$$

for some quasi-order Z and Z-permissible subset H of N.

Structure of Multiple Network Semigroups

In some cases, networks define semigroups with certain structural features.

An element s in a semigroup S possesses a (generalized) *inverse* if there exists some $t \in S$ such that $sts = s$ and $tst = t$. The element t is termed the inverse of s. If every $s \in S$ possesses an inverse, then S is termed an *inverse semigroup*. If, in addition, S possesses an *identity* element, that is, an element e such that for any $s \in S$, there exists some $t \in S$ such that $ts = e$, then S is a *group*. The element t is termed a *(group-)inverse* of s. An element s in a semigroup S is *regular* if there exists some $t \in S$ such that $sts = s$. The semigroup S is *regular* if each of its elements is regular.

It is clear from these definitions that any group is also an inverse semigroup and that any inverse semigroup is also regular. The results below describe some of the conditions under which mutiple networks give rise to regular semigroups, inverse semigroups and groups.

We first describe the conditions under which S is a group [28,42]. Let X be a network relation having a group-inverse Y, that is, a relation Y such that $XY' = I$ where I is the identity relation (for which $I(k,l) = 1$ if $k = l$; $I(k,l) = 0$ otherwise). Then it follows that (i) $XX' = I = X'X$ and $Y = X$ (that is, X' is a two-sided group-inverse for X and is unique); and (ii) X is a permutation relation (that is a relation in which, for each node k, there is exactly one node l for which $X(k,l) = 1$ and exactly one node j for which $X(j,k) = 1$). Hence if each relation X_i of a multiple network is a permutation relation and X'_i is included in S for each i, then the semigroup S is a group.

An important class of social relations which are constructed as permutation relations are "marriage class systems". These systems were proposed by White as models for kinship relations in certain Australian aboriginal tribes [48]. In such systems, persons are uniquely and permanently assigned to a *clan* and clans are related to one another by marriage and descent rules which can be represented as permutation relations. The marriage relation specifies the clan containing the wives of the male members of each clan, while the descent relation identifies the clan containing the children of male members of each clan. It follows that marriage class systems give rise to groups that represent the kinship structures of the societies concerned.

Another structure arising in the description of kinship systems is the inverse semigroup [9]. One can verify whether a network semigroup is an inverse semigroup by determining whether each of its elements has an inverse in S. The following is a characterization of conditions under which a relation possesses an inverse [24].

A network relation X has a *Thierrin–Vagner inverse* if there exists a relation Y satisfying $XYX = X$ and $YXY = Y$. Clearly, X has a Thierrin–Vagner inverse if and only if X possesses an inverse in S. If X is a regular element of S, so that $XYX = X$ for some Y in S, then YXY is a Thierrin–Vagner inverse for X. Thus determining whether X is regular will reveal whether X has an inverse. The regularity of X can be ascertained as follows.

The network relation Y is a *subinverse* of X if $XYX \leq X$. The set of subinverses of X is closed under union (that is, if Y and Z are subinverses of X then so is the relation $Y \cup Z$); hence there is a largest subinverse X^* for any relation X. The largest subinverse X^* of a relation X can be calculated from X according to $X^* = (X^T X^c X^T)^c$ where X^c denotes the complement of X. It therefore follows that a network X is regular if and only if $X \leq X(X^T X^c X^T)^c X$, and that the largest Thierrin–Vagner inverse of a regular relation X is $X^* X X^*$ [24,44].

The Algebraic Structure of Local Role Algebras

Less is known of the general algebraic properties of local role algebras. It is easily seen, however, that any orderings among relations present in the semigroup of a multiple network are also present in the local role algebra of each actor in the network. As a result, a number of the algebraic properties of local role algebras are inherited from the "parent" network semigroup. For example, if $s \leq t$ in the partially ordered semigroup S constructed from the composition operation, then $s \leq t$ in the local role algebra of every node in the network. Indeed, the converse is readily seen to also hold, so that $s \leq t$ in S if and only if $s \leq t$ in the local role algebra of every node in the network.

Algebraic Analysis

In order to describe an algebraic representation, general procedures for analyzing finite algebraic representations into simpler, independent components have been developed [20,36,37,39]. These procedures have been applied to the decomposition of lattice representations [20], partially ordered semigroups and local role algebras [34,35,36]. The procedures extract maximally independent components of a given algebraic representation and lead to a detailed structural analysis.

The account of these procedures for decomposing a finite algebra into simpler components is based on definitions adapted from [2,18]. Clearly, any decomposition procedure depends on the synthesis rules by which the components are assumed to be combined to produce the algebra. Two important synthesis rules are the direct and subdirect product.

Let $A_1 = [S_1, F, \leq]$, $A_2 = [S_2, F, \leq]$, $\ldots, A_q = [S_q, F, \leq]$ be a collection of similar algebras. The *direct product* $A_1 \times A_2 \times \cdots \times A_q$ of $A_1, A_2, \ldots, A_q$ is the algebra comprising the set $S_1 \times S_2 \times \cdots \times S_q$ and the operations $f_\alpha \in F$ given by

$$f_\alpha([x_1, x_2, \ldots, x_q], \ldots, [z_1, z_2, \ldots, z_q]) \\ = [f_\alpha(x_1, \ldots, z_1), \ldots, f_\alpha(x_q, \ldots, z_q)]\ .$$

The partial order for $A_1 \times A_2 \times \cdots \times A_q$ is given by

$$[x_1, x_2, \ldots, x_q] \leq [z_1, z_2, \ldots, z_q]$$

if and only if

$$x_1 \leq z_1, x_2 \leq z_2\,, \ldots, \text{ and } x_q \leq z_q\ .$$

An algebra $B = [T, F, \leq]$ is a *subalgebra* of $A = [S, F, \leq]$ if T is a subset of S (possibly empty) which is closed under the operations of F (that is, $f_\alpha(x_1, x_2, \ldots, x_{n(\alpha)}) \in T$, for any $x_1, x_2, \ldots, x_{n(\alpha)} \in T$, and any $f_\alpha \in F$); in addition, $x \leq y$ in B if and only if $x \leq y$ in A, for any $x, y \in T$. A subalgebra $C = [S, F, \leq]$ of a direct product $A_1 \times A_2 \times \cdots \times A_q$ of similar algebras $A_i = [S_i, F, \leq]\,(i = 1, 2, \ldots, q)$ is a *subdirect product* of $A_1, A_2, \ldots, A_q$ if for each $x_i \in S_i$, there exists an element $c \in S$ having x_i as its component in A_i.

In the case of both the direct and subdirect product of algebras, the operations in the composite algebra are performed as the conjunction of their independent operation in the component algebras. The difference between the two constructions lies in the set on which these operations are defined. For the direct product, the appropriate set is the full Cartesian product of elements from each component algebra; in the subdirect product case, the appropriate set is simply a subset of the full Cartesian product with the additional condition that each element in each component algebra appears in *some* element of the subset.

Some well-known theorems in universal algebra establish that the existence and nature of direct and subdirect representations of an algebra are determined by the lattice of homomorphisms of the algebra, or, equivalently, by a lattice of relations on the algebra associated with its homomorphisms.

An (*isotone*) *homomorphism* from an algebra $A = [S, F, \leq]$ onto a similar algebra $B = [T, F, \leq]$ is a mapping $\phi\colon S \to T$ such that,

(i) for all $f_\alpha \in F$ and $x_i \in S$,

$$\phi(f_\alpha(x_1, x_2, \ldots, x_{n(\alpha)})) \\ = f_\alpha(\phi(x_1), \phi(x_2), \ldots, \phi(x_{n(\alpha)}))\,; \quad \text{and}$$

(ii) $x \leq y$ in S implies $\phi(x) \leq \phi(y)$ in T.

The algebra B is termed a (*homomorphic*) *image* of A, and we write $B = \phi(A)$.

Further, each homomorphism ϕ from $A = [S, F, \leq]$ onto $B = [T, F, \leq]$ has a corresponding binary relation π on S (termed here a π*-relation*) in which $(x, y) \in \pi$ if and only if $\phi(y) \leq \phi(x)$; it may readily be established that π is transitive and reflexive (and hence a quasi-order) and that the equivalence relation σ_π defined by $(x, y) \in \sigma_\pi$ if and only if $(x, y) \in \pi$ and $(y, x) \in \pi$ has the substitution property, namely:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_{n(\alpha)}, y_{n(\alpha)}) \in \sigma_\pi$$

implies

$$(f_\alpha(x_1, x_2, \ldots, x_{n(\alpha)}),\ f_\alpha(y_1, y_2, \ldots, y_{n(\alpha)})) \in \sigma_\pi\ .$$

The relation σ_π is termed a *congruence relation*. The collection of π-relations may be partially ordered by: $\pi_1 \leq \pi_2$ if $(x, y) \in \pi_1$ implies $(x, y) \in \pi_2$, for any $x, y \in S$. Under this partial ordering, the relations π form a lattice $L_\pi(A)$, the *lattice of* π*-relations* on A.

The conditions under which an algebra A may be represented as a direct or subdirect product of similar algebras $A_1, A_2, \ldots, A_r$ (that is, possesses *direct* or *subdirect representations*) can be described as follows.

Let $X = \{\pi_1, \pi_2, \ldots, \pi_q\}$ be a set of π relations in $L_\pi(A)$ for which

$$\pi_1 \wedge \pi_2 \wedge \cdots \wedge \pi_q = \pi_{\min}\,,$$

where $\pi_{\min}$ is the minimal element of $L_\pi(A)$. Then A may be represented as a subdirect product of the algebras $\phi_1(A), \phi_2(A), \ldots, \phi_q(A)$, where $\phi_i(A)$ is the image of A under the homomorphism ϕ_i corresponding to the relation π_i $(i = 1, 2, \ldots, q)$. If, in addition, for all i,

(a) $\pi_1 \wedge \pi_2 \wedge \cdots \wedge \pi_{i-1}$ and π_i are *permutable* (i. e., $(\pi_1 \wedge \pi_2 \wedge \cdots \wedge \pi_{i-1})\pi_i = \pi_i(\pi_1 \wedge \pi_2 \wedge \cdots \wedge \pi_{i-1})$); and

(b) $(\pi_1 \wedge \pi_2 \wedge \cdots \wedge \pi_{i-1}) \vee \pi_i = \pi_{\max}$, where $\pi_{\max}$ is the maximal element of $L_\pi(A)$, then A may be represented as the direct product of the algebras $\phi_1(A), \phi_2(A), \ldots, \phi_q(A)$ [2].

An algebra which possesses no nontrivial representation as the direct product of similar algebras is termed *directly irreducible*, while an algebra which cannot be expressed as a nontrivial subdirect product of similar algebras is *subdirectly irreducible*. Pattison and Bartlett advocated the use of subdirect representations, but sought to restrict attention to a small set of such representations rather than to the (possibly large) class of all subdirect representations of the algebra [37]; see also [7]. Their restriction entailed (a) identifying all *irredundant* subdirect representations of an algebra, and then (b) defining a partial ordering on the irredundant representations and selecting only minimal members of the resulting partially ordered set. The resulting set of minimal, irredundant subdirect representations were termed the factorizations of the algebra.

More formally, an element x of a lattice L is *meet-irreducible* if $x = x_1 \wedge x_2$ implies $x = x_1$ or $x = x_2$. A subset of lattice elements $X = \{x_1, x_2, \ldots, x_r\}$ is *irredundant* if

a) each x_i is meet-irreducible; and
b) $x_1 \wedge \cdots \wedge x_{i-1} \wedge x_{i+1} \wedge \cdots \wedge x_q, \neq x_{\min}$, for each $i = 1, 2, \ldots, q$; where $x_{\min}$ is the minimal element of the lattice L.

The collection of all irredundant subsets of a lattice L may be partially ordered by the relation:

$X \leq Y$ iff, for each $j = 1, 2, \ldots, p$, there exists some i $(i = 1, 2, \ldots, q)$ such that $y_j \leq x_i$ in L, where $X = \{x_1, x_2, \ldots, x_q\}$ and $Y = \{y_1, y_2, \ldots, y_p\}$.

A *factorization* of an algebra A is then the subdirect representation corresponding to any minimal, irredundant subset of elements of $L_\pi(A)$ whose meet is the minimal relation $\pi_{\min}$.

An Algorithm for Factorization

A general algorithm for constructing the set of all factorizations of an algebra has also been developed [37]. The algorithm operates on a subset of relations in the lattice $L_\pi(A)$ and conducts a reasonably efficient search for subsets of π relations which satisfy the conditions of the factorization definition. In particular, the algorithm constructs the set of factorizations of an algebra A from the atoms of its lattice $L_\pi(A)$ of π-relations and their maximal meet-complements.

An *atom* of a lattice L is an element that covers $x_{\min}$. A *meet-complement* of an element $x \in L$ is an element x^* such that $x^* > x_{\min}$ and $x \wedge x^* = x_{\min}$. A meet-complement x^* of x is *maximal* if x has no other meet-complement z such that $z > x^*$.

Let $Z = \{z_1, z_2, \ldots, z_a\}$ be the set of atoms of $L_\pi(A)$. Define the collection of sets of the form $\{z_1^*, z_2^*, \ldots, z_a^*\}$ where z_i^* is a maximal meet-complement of z_i. Then it can be shown that any factorization of A is a subset of such a set [36]. This result is the basis for the algorithm for finding factorizations. Under certain circumstances, we can write down an explicit expression for π-relations in the factorization of A.

Let A be an algebra with π-relation lattice $L_\pi(A)$. Let π_{ab} be the least π-relation in which $a \geq b$ for elements $a, b \in A$; we term π_{ab} the π-relation generated by the ordering $a \geq b$.

Let z be an atom of $L_\pi(A)$; define the relation $\pi(z) = \{(a, b) : \pi_{ab} \wedge z = \pi_{\min}\}$. If z has a unique maximal meet-complement, then $\pi(z)$ is a π-relation and is a unique maximal meet-complement of z. Conversely, if $\pi(z)$ is a π-relation, then it is the unique maximal meet-complement of z. Further, if $Z = \{z_1, z_2, \ldots, z_a\}$ is the set of atoms of $L_\pi(A)$ and if, for each $z_i \in Z$, z_i has a unique maximal meet-complement $\pi(z_i)$, then the factorization of A is unique and corresponds to the π-relations $\{\pi(z_1), \pi(z_2), \ldots, \pi(z_a)\}$ [36]. In other words, if the factorization is unique, it can be identified immediately from the maximal meet-complements of the atoms of $L_\pi(A)$.

Under some circumstances, we can make more general claims about the uniqueness of factorizations. Not surprisingly, the structure of the lattice $L_\pi(A)$ plays a major part. It is known, for instance, that if the lattice $L_\pi(A)$ is distributive, then A possesses a unique irredundant subdirect representation, and hence a unique factorization. In particular, since the lattice of π-relations for a lattice is necessarily distributive, any finite lattice has a unique factorization. If the lattice $L_\pi(A)$ is modular, then any irredundant subdirect representation of A has the same number of components [2].

Factorization for Partial Algebras

Analogous decompositions can be described for partial algebras. An (*isotone*) *homomorphism* from a partial algebra $A = [S, F, \leq]$ onto a similar partial algebra $B = [T, F, \leq]$ is a mapping $\phi \colon S \rightarrow T$ such that,

(i) for all $f_\alpha \in F$ and $x_i \in S$,

$$\begin{aligned}\phi(f_\alpha(x_1, x_2, \ldots, x_{n(\alpha)}))\\ = f_\alpha(\phi(x_1), \phi(x_2), \ldots, \phi(x_{n(\alpha)}))\end{aligned}$$

whenever the operations on both sides of the expression are defined; and

(ii) $x \leq y$ in S implies $\phi(x) \leq \phi(y)$ in T.

The algebra B is termed a (*homomorphic*) *image* of A, and we write $B = \phi(A)$.

Further, each homomorphism ϕ from $A = [S, F, \leq]$ onto $B = [T, F, \leq]$ has a corresponding binary relation π on S (termed here a π*-relation*) in which $(x, y) \in \pi$ if and only if $\phi(y) \leq \phi(x)$; it may readily be established that π is transitive and reflexive (so that π is a quasi-order). Further, the equivalence relation σ_π defined by $(x, y) \in \sigma_\pi$ if and only if $(x, y) \in \pi$ and $(y, x) \in \pi$ has the substitution property, namely:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_{n(\alpha)}, y_{n(\alpha)}) \in \sigma_\pi$$

implies

$$(f_\alpha(x_1, x_2, \ldots, x_{n(\alpha)})\,,\ f_\alpha(y_1, y_2, \ldots, y_{n(\alpha)})) \in \sigma_\pi$$

whenever the latter term is defined. The relation σ_π is often termed a *congruence relation*. The collection of all π-relations form a lattice $L_\pi(A)$, the *lattice of* π*-relations* on A, under the partial ordering: $\pi_1 \leq \pi_2$ if $(x, y) \in \pi_1$ implies $(x, y) \in \pi_2$, for any $x, y \in S$.

In the case of partial algebras, it is convenient to express the algebra as an intersection of the π-relations corresponding to maximally independent homomorphic images of the algebra. Recall that the minimal element $\pi_{\min}$ of the lattice $L_\pi(A)$ corresponds to the algebra A itself. Thus, any expression of $\pi_{\min}$ in the form

$$\pi_1 \wedge \pi_2 \wedge \cdots \wedge \pi_q = \pi_{\min}$$

expresses A as the intersection of π-relations corresponding to homomorphic images of A. Any such expression that involves π-relations $\{\pi_i\,;\ i = 1, 2, \ldots, q\}$ satisfying the conditions of the factorization definition yields an expression in terms of a set of simple and maximally independent images of the algebra. Further, the same algorithm for constructing factorizations of an algebra that was described earlier may be used to find expressions of this type.

Algebraic and Network Mappings

The semigroup of a multiple network records the orderings and equations among different types of labelled paths in the network. A natural question to arise is how properties of a multiple network constrain the relational structure of its semigroup. We therefore review some general conditions under which the semigroup of one network is a homomorphic image of the semigroup of another.

We define two multiple networks to possess the same structure if their semigroups are isomorphic, that is, if the network relations are in one-to-one correspondence and they have identical right multiplication tables and partial orders.

One strategy for investigating whether networks have similar structures is to determine the conditions under which network homomorphisms induce homomorphisms of the network semigroup [4]. A *network homomorphism* from a multiple network $\{X_1, X_2, \ldots, X_r\}$ on actor set N to a multiple network $\{Y_1, Y_2, \ldots, Y_r\}$ on actor set M is a mapping ψ from N onto M such that: (a) $X_m(k, l) = 1$ implies $Y_m(\psi(k), \psi(l)) = 1$, for any k, l, m; and (b) $Y_m(i, j) = 1$ for some i, j implies that $X_m(k, l) = 1$, for some k, l such that $\psi(k) = i$ and $\psi(l) = j$. The network on M is termed the *image* of the network on N under the mapping ψ.

The mapping ψ satisfies the *structural equivalence* condition if for any m, and for any $k, l \in N$, $\psi(k) = \psi(l)$ if and only if:

- $X_m(k, j) = 1$ iff $X_m(l, j) = 1$ for any $j \in N$; and
- $X_m(j, k) = 1$ iff $X_m(j, l) = 1$ for any $j \in N$.

Lorrain and White observed that if two multiple networks are related by a network homomorphism satisfying the structural equivalence condition, then their semigroups are isomorphic and we may argue that they possess the *same* relational structure [27]. More generally, we can ask whether a homomorphism between two networks induces a homomorphism between their semigroups. It is readily seen that a homomorphism is not always guaranteed.

Network homomorphisms induced by certain blockmodels lead to semigroup homomorphisms (see the entry on ► Positional Analysis and Blockmodeling). More general conditions under which a semigroup homomorhism is guaranteed were established by Kim and Roush [25].

A network homomorphism ψ satisfies Kim and Roush's *condition* G_i if the following holds for any pair of equivalence classes ρ_1 and ρ_2 on N induced by the mapping ψ (so that $k, l \in \rho_h$ for some h iff $\psi(k) = \psi(l)$). Let the number of elements in ρ_1 and ρ_2 be n_1 and n_2, respectively. Let D be any subset of ρ_1 of i elements or, if $i > n_1$, let $D = \rho_1$. Then, for any $m = 1, 2, \ldots, r$, the set $\{l : l \in \rho_2 \text{ and } X_m(k, l) = 1 \text{ for some } k \in D\}$ has at least $\min(i, n_2)$ elements. The condition G_1 is also known as the *outdegree condition* and G_n is also termed the *indegree condition*. A network homomorphism that satisfies both G_1 and G_n is termed *regular*.

Kim and Roush demonstrated that if ψ is a network homomorphism from one multiple network onto another that satisfies the condition G_i, then there is a homomorphism mapping the semigroup of the first onto the semigroup of the second.

A more general condition combines the condition G_i with what Pattison termed the *central representatives condition* [34]. Let ψ be a network homomorphism. Then ψ satisfies the condition G_{im} if, for each class ρ of elements of N induced by ψ,

- there exists a *central subset* C of ρ such that for any X_m
  - $X_m(k, l) = 1$ for some $k \in \rho$ implies $X_m(k^*, l) = 1$ for some $k^* \in C$, and
  - $X_m(l, k) = 1$ for some $k \in \rho$ implies $X_m(l, k^*) = 1$ for some $k^* \in C$; and
- if C^* denotes the union of central subsets C, then the central subsets C satisfy Kim and Roush's condition G_i on the network defined on C^*.

If each central subset C comprises a single actor, then the condition is equivalent to Pattison's central representatives condition, while if each central subset C comprises the whole of the equivalence class on N induced by ψ, it is equivalent to the condition G_i. Kim and Roush showed that if one network can be mapped onto another by a network homomorphism satisfying the condition G_{im}, then there is a homomorphism from the semigroup of the first to the semigroup of the second [25]. The condition G_{im} is the most general condition known that guarantees the existence of such a homomorphism.

Comparing Network Semigroups

Two network semigroups S_1 and S_2 are strictly comparable only if $S_1 \leq S_2$ or $S_2 \leq S_1$. Loosely speaking, though, they are "similar" if they share many homomorphic images. Boorman and White proposed that the largest shared homomorphic image (the so-called *joint homomorphism*) is a useful construction for comparing network semigroups [5]. Bonacich and McConaghy, on the other hand, argued that the smallest semigroup containing each of S_1 and S_2 (the *common structure semigroup*) was a more appropriate representative of common semigroup structure [3,32]. The resolution of the problem of finding a representative of common structure depends on how a semigroup is viewed [36]. If a semigroup is seen as a list of the homomorphic images that it admits, then the joint homomorphism corresponds to the list of shared features, while if it seen as a collection of semigroup equations and orderings, then the common structure semigroup serves as a representation of shared structure. The common structure semigroup of two network semigroups is the semigroup of the disjoint union of the underlying networks.

Both the joint homomorphic image and common structure semigroup constructions have been shown to be useful in particular applications, the former in identifying common reductions of the networks giving rise to identical semigroups, and the latter in identifying shared cultural forms (such as "the friend of a friend is always an associate").

Linking Algebraic and Network Homomorphisms

The factorization of an algebra identifies relatively independent components of the algebra generated by constituents of the data array under the selected operation. It is natural to ask whether this analysis induces a corresponding decomposition of the network itself. That is, if an algebra A has a homomorphic image B (such as one of the factors of A), is it possible to find one or more homorphisms of the network that is consistent with the congruence relation σ for the homomorphism and that generates the algebra B? If so, we can argue that there is an association between such a reduction of the network and the homomorphic image.

In the case of the lattice of an affiliation network, such a reduction can always be uniquely found [20]. For semigroups and role algebras, there is no guarantee that such a reduction can be found, but it is possible to identify the smallest homomorphic image of the network whose algebra contained B as a homomorphic image [34].

Examples

We illustrate a number of the algebraic constructions just described for a multiple network blockmodel and an approximation to an affiliation network.

Network Semigroup

Table 1 corresponds to a multiple network on 7 nodes with two kinds of edges, F and N. The network is a blockmodel reported by Vickers for relations of friendship and negative ties among members of a high school class in an Australian country town [46]. The questions from which the blockmodel was constructed were (a) "Who are your best friends?" and (b) "Who would you rather not have as a friend?".

The semigroup generated by the multiple network in Table 1 has the right multiplication table and partial order shown in Table 2. The process of constructing the semigroup generates the sets $W_1 = \{F, N\}$, $W_2 = \{F^2, FN, NF, N^2\}$, $W_3 = \{F^3, FNF, FN^2, NF^2, N^2F, N^3\}$, $W_4 =$

Social Networks, Algebraic Models for, Table 1
A multiple network on 7 nodes

		Block									Block						
Relation	Block	1	2	3	4	5	6	7	Relation	Block	1	2	3	4	5	6	7
F	1	1	1	1	0	0	0	0	*N*	1	0	0	0	1	1	1	1
	2	0	1	0	0	0	0	1		2	0	0	0	1	1	1	0
	3	1	1	1	0	0	0	0		3	0	0	0	1	1	1	0
	4	0	0	0	1	0	0	0		4	0	1	1	0	0	1	1
	5	0	0	0	0	1	0	0		5	0	0	0	0	0	0	1
	6	1	0	0	0	0	1	0		6	0	0	0	1	0	0	1
	7	1	1	1	0	0	0	1		7	0	0	0	1	1	1	1

Social Networks, Algebraic Models for, Table 2
Right multiplication table and partial order for the semigroup of the multiple network of Table 1

Right multiplication table				Partial order														
		Generators			Elements													
Element	Word	*F*	*N*		1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	*F*	3	4	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	*N*	5	6	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0
3	*FF*	7	4	3	1	0	1	0	0	0	0	0	0	0	0	0	0	0
4	*FN*	8	9	4	0	1	0	1	0	0	0	0	0	0	0	0	0	0
5	*NF*	10	9	5	0	1	0	0	1	0	0	0	0	0	0	0	0	0
6	*NN*	11	12	6	0	0	0	0	0	1	0	0	0	0	0	0	0	0
7	*FFF*	7	4	7	1	0	1	0	0	0	1	0	0	0	0	0	0	0
8	*FNF*	8	9	8	0	1	0	1	1	0	0	1	0	1	0	0	0	1
9	*FNN*	13	12	9	0	0	0	0	0	1	0	0	1	0	0	0	0	0
10	*NFF*	14	9	10	0	1	0	0	1	0	0	0	0	1	0	0	0	0
11	*NNF*	11	12	11	1	0	1	0	0	1	1	0	0	0	1	0	0	0
12	*NNN*	13	12	12	0	1	0	1	0	1	0	0	1	0	0	1	0	0
13	*FNNF*	13	12	13	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14	*NFFF*	14	9	14	0	1	0	0	1	0	0	0	0	1	0	0	0	1

$\{FN^2F, NF^3\}$ and $W_5 = \emptyset$; equations include $F^2N = FN$, $NFN = FN^2$, and $F^4 = F^3$.

It is evident from Table 2 that the relations F and N both have index 3 and period 1. The relation F is regular, whereas the relation N is not.

Each of the equations of Table 2 specifies the empirical coincidence of potentially distinct labelled walks. The equation $F^2N = FN$, for example, indicates the those individuals nominated by one's friends as preferred nonfriends coincide with individuals so nominated by the friends of one's friends.

Role Algebra

The role algebra for block 4 in the network of Table 1 is presented in Table 3. Equations in the right multiplication table indicate equalities among labelled paths having block 4 as the source. For example, the equation $FF = F$ indicates that friendship paths of length 2 emanating from block 4 reach exactly the same set of blocks as direct friendship ties. Likewise, the equation $FN = N$ allows us to infer that, namely block 4 would prefer not to have as friends precisely those blocks so nominated by their friends. The partial order diagram indicates that block 4's friendship ties are a subset of those to whom block 4 has negative tie paths (block 4) of length 2.

The role algebra of block 4 has a unique factorization. The lattice of π-relations has two atoms and each atom possesses a unique maximal meet-complement; the atoms and their maximal meet-complements are shown in Table 4. These unique maximal meet-complements are associated with the role algebras shown in Table 5. In the first factor, both *FN* and *NF* ties from block 4 coincide with N ties, and *NNF* ties include all others. In the second factor, any path from block 4 whose last tie is labelled N coincides with an N tie.

Social Networks, Algebraic Models for, Table 3

The role algebra for block 4 in the network of Table 1

Right multiplication table				Partial order						
		Generators			Elements					
Element	Word	*F*	*N*		1	2	3	4	5	6
1	*F*	1	2	1	1	0	0	0	0	0
2	*N*	3	4	2	0	1	0	0	0	0
3	*NF*	3	4	3	0	1	1	0	0	0
4	*NN*	5	6	4	1	0	0	1	0	0
5	*NNF*	5	6	5	1	1	1	1	1	1
6	*NNN*	5	6	6	1	1	0	1	0	1

Social Networks, Algebraic Models for, Table 4

Atoms and unique maximal-meet complements for the block 4 role algebra

	Partial order							Partial order						
		Elements							Elements					
	Elements	1	2	3	4	5	6	Elements	1	2	3	4	5	6
Atom	1	1	0	0	0	0	0	1	1	0	0	0	0	0
	2	0	1	0	0	0	0	2	0	1	0	0	0	0
	3	0	1	1	0	0	0	3	0	1	1	0	0	0
	4	1	1	0	1	0	0	4	1	0	0	1	0	0
	5	1	1	1	1	1	1	5	1	1	1	1	1	1
	6	1	1	0	1	0	1	6	1	1	1	1	0	1
Maximal meet-complement	1	1	0	0	0	0	0	1	1	0	0	0	0	0
	2	0	1	1	0	0	0	2	1	1	0	1	0	1
	3	0	1	1	0	0	0	3	1	1	1	1	1	1
	4	1	0	0	1	0	0	4	1	1	0	1	0	1
	5	1	1	1	1	1	1	5	1	1	1	1	1	1
	6	1	1	1	1	1	1	6	1	1	0	1	0	1

Lattice of an Affiliation Network

The affiliation network in Table 6 is an approximation to the so-called "Southern Women" data [12,23]. The approximation was obtained by a method described in [52], and is akin to a form of dual clustering of the rows and columns of the original data array. The Galois lattice of the (approximate) affiliation array is presented in Fig. 1. In Fig. 1, known as the *line diagram* of the lattice, one element *s* is drawn above and connected to another element *t* if *s* covers *t*, (that is if $s \geq t$ and there is no element *u* distinct from *s* and *t* for which $s \geq u \geq t$) [51]. The meet of two elements is the greatest element lying below both elements and connected to them by a descending path. The join of two elements is the least element to which they are both connected by an ascending path.

Social Networks, Algebraic Models for, Table 5

Factors of the role algebra for block 4

Factor 1	Right multiplication table			Partial order				
		Generators			Elements			
Element	Word	*F*	*N*		1	2	3	4
1	*F*	1	2	1	1	0	0	0
2	*N*	2	3	2	0	1	0	0
3	*NN*	4	4	3	1	0	1	0
4	*NNF*	4	4	4	1	1	1	1

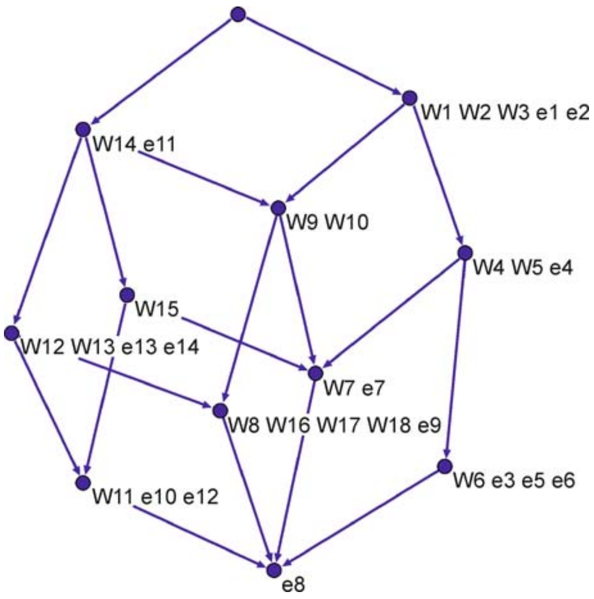
Factor 2	Right multiplication table			Partial order			
		Generators			Elements		
Element	Word	*F*	*N*		1	2	3
1	*F*	1	2	1	1	0	0
2	*N*	3	2	2	0	1	0
3	*NF*	3	2	3	0	1	1

Social Networks, Algebraic Models for, Table 6

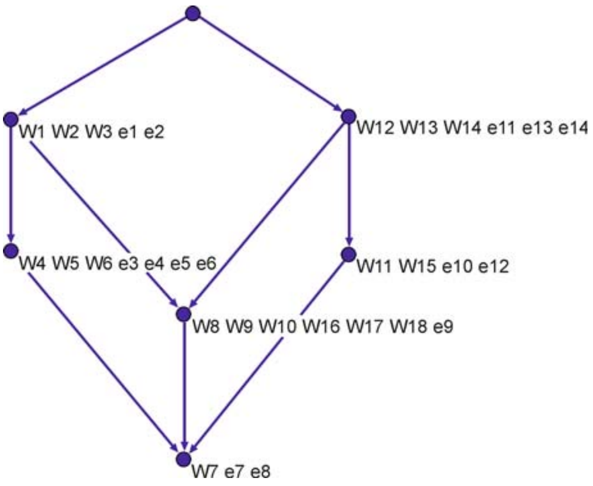
An affiliation network: an approximation to the Southern Women data; the approximation differs from the original data in the 20 underlined values

Event														
Woman	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1	1	1	1	1	1	$\underline{1}$	1	1	0	0	0	0	0
2	1	1	1	$\underline{1}$	1	1	1	1	$\underline{1}$	0	0	0	0	0
3	$\underline{1}$	1	1	1	1	1	1	1	1	0	0	0	0	0
4	$\underline{0}$	0	1	1	1	1	1	1	0	0	0	0	0	0
5	0	0	1	1	1	$\underline{1}$	1	$\underline{1}$	0	0	0	0	0	0
6	0	0	1	0	1	1	0	1	0	0	0	0	0	0
7	0	0	0	0	$\underline{0}$	$\underline{0}$	1	1	0	0	0	0	0	0
8	0	0	0	0	0	$\underline{0}$	0	1	1	0	0	0	0	0
9	0	0	0	0	$\underline{0}$	0	1	1	1	0	0	0	0	0
10	0	0	0	0	0	0	1	1	1	0	0	$\underline{0}$	0	0
11	0	0	0	0	0	0	0	1	$\underline{0}$	1	0	1	0	0
12	0	0	0	0	0	0	0	1	1	1	0	1	1	1
13	0	0	0	0	0	0	$\underline{0}$	1	1	1	0	1	1	1
14	0	0	0	0	0	$\underline{0}$	1	$\underline{1}$	1	1	1	1	1	1
15	0	0	0	0	0	0	1	1	0	1	1	1	0	0
16	0	0	0	0	0	0	0	1	1	0	0	0	0	0
17	0	0	0	0	0	0	0	$\underline{1}$	1	0	$\underline{0}$	0	0	0
18	0	0	0	0	0	0	0	$\underline{1}$	1	0	$\underline{0}$	0	0	0

The rows and columns of the affiliation array may be mapped to lattice elements in such a way that any row element (in this case, a woman) is at or above any column elements (in this case, events) with which it is affiliated; likewise, any column element (event) is at or below any affiliated row elements (women). It can be seen from Fig. 1, for example, that all women attend event e8, whereas only women W1, W2 and W3 attend events e1 and e2.



Social Networks, Algebraic Models for, Figure 1
The line diagram of the Galois lattice of Table 4

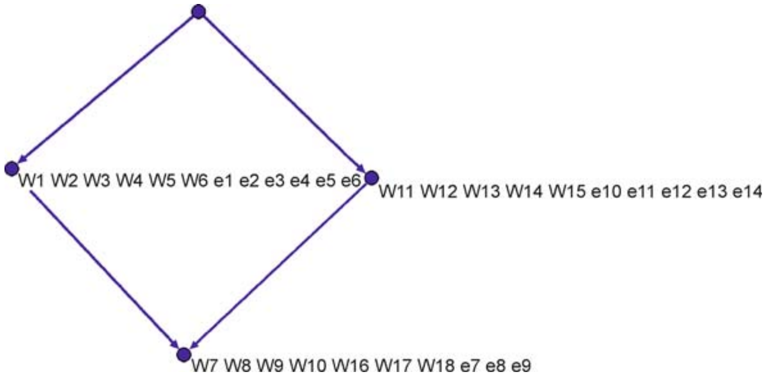


Social Networks, Algebraic Models for, Figure 2
Largest homomorphic image of the lattice of Fig. 1

The lattice displayed in Fig. 2 is a maximal homomorphic image of the lattice of Fig. 1 and provides a simplification of the lattice. In this diagram, a division of the women in the network into three ordered clusters is apparent. One cluster comprises the women W1 to W6, with women W4, W5 and W6 attending only a subset of the events attended by women W1, W2 and W3. A second cluster comprises women W11 to W15, with women W11 and W15 attending only a subset of the events attended by women W12, W13 and W14. The third cluster of women comprises W7, W8, W16, W17 and W18 and is again divided into two sub-clusters, one of whom (W7) attends a subset of the events attended by the other (W8, W16, W17, W18). It is also clear from the lattice of Fig. 2 that the first two clusters of women attend a mixture of distinct events (e1 to e6 in the case of the first cluster, and e10 to e14 in the case of the second) as well as common events (e7 to e9), whereas the third cluster of women attend only those events also attended by women in the first two clusters (e7 to e9). This structural feature of the affiliation network is drawn out very clearly in the unique maximal homomorphic image of the lattice of Fig. 2, shown in Fig. 3.

Future Directions

The algebraic constructions described above provide exact and detailed representations of structural relationships


W1 W2 W3 W4 W5 W6 e1 e2 e3 e4 e5 e6
W11 W12 W13 W14 W15 e10 e11 e12 e13 e14
W7 W8 W9 W10 W16 W17 W18 e7 e8 e9

Social Networks, Algebraic Models for, Figure 3
Largest homomorphic image of the lattice of Fig. 2

among network constituents. Of course, this level of detail assumes that the relational observations comprising the network are accurate, as no allowance is made for variability or error. This is often a tenuous assumption, and two important directions for further development utilize the structural sensitivities of algebraic representations in the presence of network tie variability.

The first direction is to use statistical criteria to generate (partial) algebraic structures that summarize key structural features of a network [1,8,39,40]. Instead of the generalized Axiom of Quality introduced earlier, an *Approximate Axiom of Quality* is proposed in which $u \leq v$ in W whenever there is "sufficient evidence" that the relation $X_u \leq X_v$ holds. Such an approach can lead to a theoretically-guided and structurally-focused form of exploratory data analysis for multiple networks. Theoretical guidance comes from choice of operations in the set F, while structural focus resides in the (partial) algebra to which the approach gives rise.

The second direction is to understand how exact algebraic representations can emerge as special, so-called *degenerate*, cases of statistical models such as exponential random graph models for networks (see the entry on ▶ Social Networks, Exponential Random Graph ($\boldsymbol{p}^*$) Models for *for Networks*). An exponential random graph model defines a probability distribution on the set of all networks on a node set N as a function of some parameter vector. For any parameter vector, a subset of networks have what can be defined as minimum "energy". As the parameter vector is scaled by an increasingly large multiplier, the probability associated with any network whose energy exceeds the minimum tends to zero. The minimum energy networks can therefore be seen as highly constrained or "frozen" structural forms associated with the stochastic model [21,22,41]. The structure of these forms will often warrant algebraic analysis of the type described earlier; in addition, we can understand the non-frozen models as stochastic generalizations of these structural forms.

Bibliography

1. Barnes G, Cerrito P, Levi I (1998) A mathematical model for interpersonal relationships in social networks. Soc Netw 20:179–196
2. Birkhoff G (1963) Lattice theory, 3rd edn. American Mathematical Society, Providence
3. Bonacich P (1980) The common structure semigroup, a replacement for the Boorman and White joint reduction. Am J Soc 86:159–166
4. Bonacich P (1983) Representations for homomorphisms. Soc Netw 5:173–192
5. Boorman SA, White HC (1976) Social structures from multiple networks II. Role structures. Am J Soc 81:1384–1446
6. Boyd JP (1969) The algebra of group kinship. J Math Psychol 6:139–167
7. Boyd JP (1989) Social semigroups and Green relations. In: Freeman LC, White DR, Romney AK (eds) Research methods in social network analysis. George Mason University Press, Fairfax, pp 215–254
8. Boyd JP (1990) Social semigroups: A unified theory of scaling and blockmodelling as applied to social networks. George Mason University Press, Fairfax
9. Boyd JP, Haehl JH, Sailer LD (1989) Kinship systems and inverse semigroups. J Math Soc 2:37–61
10. Breiger RL (1974) The duality of persons and groups. Soc Forces 53:181–190
11. Breiger RL, Pattison PE (1986) Cumulated social roles: the duality of persons and their algebras. Soc Netw 8:215–256
12. Davis A, Gardner B, Gardner M (1941) Deep south. Chicago University Press, Chicago
13. Doreian P (1981) Polyhedral dynamics and conflict mobilization in social networks. Soc Netw 3:107–116
14. Duquenne V (1996) On lattice approximations: syntactic aspects. Soc Netw 18:189–200
15. Freeman LC (1996) Cliques, Galois lattices, and the structure of human social groups. Soc Netw 18:173–87
16. Freeman LC, White DR (1993) Using Galois lattices to represent network data. In: Marsden P (ed) Sociological methodology. American Sociological Association, Washington, pp 127–146
17. Friedell M (1967) Organizations as semilattices. Am Soc Rev 32:46–54
18. Fuchs L (1963) Partially ordered algebraic systems. Pergamon, Oxford
19. Ganter B, Wille R (1989) Conceptual scaling. In: Roberts F (ed) Applications of combinatorics and graph theory in the biological and social sciences. Springer, New York, pp 139–167
20. Ganter B, Wille R (1999) Formal concept analysis: mathematical foundations. Springer, New York
21. Grenander U (1993) General pattern theory. Oxford University Press, Oxford
22. Handcock MS (2003) Statistical models for social networks. In: Breiger RL, Carley KM, Pattison PE (eds) Dynamic social network modeling and analysis. National Academies Press, Washington
23. Homans G (1951) The human group. Routledge & Kegan Paul, London
24. Kim KH (1982) Boolean matrix theory and applications. Dekker, New York
25. Kim KH, Roush FW (1984) Group relationships and homomorphisms of Boolean matrix semigroups. J Math Psychol 28:448–452
26. Lorrain F (1975) Reseaux sociaux et classifications sociales. Hermann, Paris
27. Lorrain F, White HC (1971) Structural equivalence of individuals in social networks. J Math Soc 1:49–80
28. Luce RD (1956) A note on Boolean matrix theory. Proc Am Math Soc 3:382–388
29. Mandel M (1983) Local roles and social networks. Am Soc Rev 48:376–386
30. Martin JL (2002) Some algebraic structures for diffusion in social networks. J Math Soc 26:123–146

31. Martin JL (2006) Jointness and duality in algebraic approaches to dichotomous data. Soc Method Res 35:159–192
32. McConaghy M (1981) The common role structure: improved blockmodelling methods applied to two communities' elites. Soc Method Res 9:267–285
33. Mische A, Pattison PE (2000) Composing a civic arena: Publics, projects and social settings. Poetics 27:163–194
34. Pattison PE (1982) The analysis of semigroups of multirelational systems. J Math Psychol 25:87–118
35. Pattison PE (1989) Mathematical models for local social networks. In: Keats J, Taft R, Heath R, Lovibond S (eds) Mathematical and theoretical systems. North Holland, Amsterdam
36. Pattison PE (1993) Algebraic models for social networks. Cambridge University Press, New York
37. Pattison PE, Bartlett WK (1982) A factorization procedure for finite algebras. J Math Psychol 25:51–81
38. Pattison PE, Breiger RL (2002) Lattices and dimensional representations: matrix decompositions and ordering structures. Soc Netw 24:423–444
39. Pattison PE, Wasserman S (1995) Constructing algebraic models for local social networks using statistical methods. J Math Psychol 39:57–72
40. Pattison PE, Wasserman S, Robins G, Kanfer A (2000) Statistical evaluation of algebraic constraints for social networks. J Math Psychol 44:536–568
41. Robins G, Woolcock J, Pattison P (2005) Small and other worlds: Global network structures from local processes. Am J Sociol 110:894–936
42. Rutherford D (1963) Inverses of Boolean matrices. Proc Glasgow Math Assoc 6:49–53
43. Schein B (1970) A construction for idempotent binary relations. Proc Jpn Acad 46:246–247
44. Schein B (1976) Regular elements of the semigroup of all binary relations. Semigr Forum 13:95–102
45. Schweizer T (1993) The dual ordering of actors and possessions. Curr Anthropol 34:469–483
46. Vickers M (1981) Relational analysis: an applied evaluation. MSc Thesis, University of Melbourne, Melbourne
47. White DR, Duquenne V (1996) Special issue on social networks and discrete structure analysis. Soc Netw 18:169–318
48. White HC (1963) An anatomy of kinship. Prentice-Hall, Englewood Cliffs
49. White HC (1992) Identity and control. University of Chicago Press, Chicago
50. White HC, Boorman SA, Breiger RL (1976) Social structure from multiple networks: I. Blockmodels of roles and positions. Am J Sociol 81:730–780
51. Wille R (1984) Line diagrams of hierarchical concept systems. Intern Classif 11:77–86
52. Wilson S, Bladin P, Saling M, Pattison P (2005) Characterizing psychosocial outcome trajectories following seizure surgery. Epilepsy Behav 6:570–580
53. Winship C, Mandel M (1983) Roles and positions: a critique and extension of the blockmodelling approach. In: Leinhardt S (ed) Sociological methodology. Jossey-Bass, San Francisco, pp 314–344
54. Wu L (1983) Local blockmodel algebras for analyzing social networks. In: Leinhardt S (ed) Sociological methodology. Jossey-Bass, San Francisco, pp 272–313

Social Networks, Diffusion Processes in

THOMAS W. VALENTE
Department of Preventive Medicine, School of Medicine,
University of Southern California, Alhambra, USA

Article Outline

Glossary

Adoption A person's change in behavior.

Agent based models Creation of hypothetical (sometimes prototypical) network structures and the simulation of diffusion within those structures.

Homophily Tendency for people to be connected to others like themselves.

Incidence The percent of new adopters at each time period.

Internal versus external influence Internal influence posits that adoption is driven by person-to-person persuasion whereas external influence posits it is driven by sources outside the network such as mass media.

Event history analysis Transformation of data to represent person-time observations.

Network exposure The degree of behavioral adoption in each person's network neighborhood.

Network threshold The number or percent of adopters in a person's neighborhood necessary for a person to adopt the innovation.

Rate The speed of diffusion.

Prevalence The cumulative percent of adopters in the population.

Weight matrix Any N×N matrix representing potential distances or similarities that models potential pathways for adoption influence (e. g., a structural equivalence matrix derived from an adjacency matrix to model influence of structural equivalence relations on adoption).

Definition of the Subject

Diffusion of innovations is the study of how new ideas and behaviors spread within a population, both within and between communities. Diffusion research spans many disciplines since understanding how new ideas and behaviors spread is an important aspect in many different fields of study. For example, marketers are often interested in how and why people change their purchasing habits and sociologists concerned with how people come to accept new norms. The first empirical investigations date back 100 years, while contemporary studies mark the beginning to the 1943 study of diffusion of hybrid seed corn in Iowa. A subset of the hundreds of empirical diffusion studies have focused specifically on the collection and analysis of social network data combined with data on when individuals adopted selected behaviors.

Introduction

Diffusion of innovations theory attempts to explain how new ideas and practices spread within and between communities. The theory has its roots in anthropology, economics, geography, sociology, marketing, mathematics, among other disciplines [1,2,3,4], and has in some ways been adapted from epidemiology [5,6]. The premise, confirmed by empirical research, is that new ideas and practices spread through interpersonal contacts largely consisting of interpersonal communication [1,7,8,9,10,11].

In their 1943 pioneering study, Ryan and Gross [10] laid the groundwork for the diffusion paradigm by showing that, among other things, social factors rather than economic ones were important influences on adoption [11]. Hundreds of diffusion studies were conducted in the 1950s and early 1960s to examine the diffusion process in more detail across a variety of settings [1]. Many studies sought to understand how information created in government or otherwise sponsored programs could be disseminated more effectively. Diffusion research peaked in the early 1960s, but has been reinvigorated recently with the advent of more sophisticated network models and technology making it possible to study the diffusion process more explicitly.

Most diffusion studies focus on trying to understand the factors that lead some members of a population to adopt a new idea, while others do not. Further, studies try to understand why some people adopt the behavior early while others wait a substantial amount of time before accepting the new practice. For example, Ryan and Gross [10] wanted to know why some farmers purchased hybrid seed corn almost immediately upon its availability while others waited until almost all the farmers in the area purchased it before they were willing to do so. Similarly, Coleman and others [12] wanted to know why some physicians began prescribing tetracycline as soon as it was available, while others waited until most physicians prescribed it before they were willing to do so.

This chapter describes a variety of mathematical and network models used to study the diffusion of these and other innovations. The Coleman and others [12] study provided a conceptual leap from other diffusion studies by explicitly measuring who talked to whom within the community about the innovation. Rogers [13] also collected network data to study the diffusion of weed spray in Iowa in his dissertation. Burt [14] unearthed the data and made it available to the network community so that scholars could debate various models used to describe the network diffusion process. Although having data has been useful for clarifying diffusion models, the limitations of these data and this study, make it a poor choice for studying adoption behavior. Rather, scholars should have focused on collecting better data, or reanalyzing diffusion network data in which contagion was more likely.

This chapter chronicles the development of network diffusion models and indicates where such progress is being made. I first present macro models used to estimate the speed of diffusion and, with the Bass [15] model, to estimate rates of innovation and imitation. Next, spatial autocorrelation is presented which is used to estimate the degree contiguous nodes adopt innovations. Spatial autocorrelation led to the network autocorrelation model which is presented statically (cross-sectional data only) and then with one time lag. I then discuss event history analysis applications of network autocorrelation and its extension by including time-based network interaction terms. Throughout the chapter, I attempt to provide a review of recent research conducted in a variety of domains, but mostly drawn from the public health field.

Macro Models

One consistent finding of diffusion research has been that the cumulative pattern of diffusion follows a growth pattern approximated by a simple one-parameter logistic function such as:

$$y_t = b_0 + \frac{1}{1 + \mathrm{e}^{-b_1 t}} , \qquad (1)$$

where y is the proportion of adopters, b_0 the y intercept, t is time, and b_1 the rate parameter to be estimated. This simple model can be used to compare growth rates for various innovations, but is extremely limited in its applicability. A considerable improvement was advanced by Bass [15]

Social Networks, Diffusion Processes in, Table 1
Diffusion Rate Parameter Estimates and Moran's *I* Estimates for Two Data Sets

	Medical Innovation	Cameroon Tontine 1 Simulation
One Parameter Model		
Coefficient (95% Confidence Intervals)	0.23 (–.053–0.51)	0.06 (.01–0.12)
N	17	50
R^2	0.76	0.71
Two Parameter (Bass) Model		
Innovation Coefficient (95% C. I.)	–0.43(–0.83–0.03)	–0.20 (–0.30–0.09)
Imitation Coefficient (95% C. I.)	4.09 (3.05–5.12)	2.96 (2.58–3.34)
N	16	49
R^2	0.89	0.89
Moran's *I*	–.13	–.08
z-score	–6.73	–7.80

and many others [16,17,18] by creating a two parameter model:

$$y_t = b_0 + (b_1 - b_0)Y_{t-1} - b_1(Y_{t-1})^2 \,, \tag{2}$$

where y is the proportion of adopters, b_0 a rate parameter for innovation, and b_1 a rate parameter for imitation (the degree of adoption due to prior adopters). The Bass model incorporates the percentage adopters at each time point and thus makes a better estimate of the growth attributable to personal network persuasion. The mathematical model in Eq. 2 can be used to: (1) forecast expected levels of diffusion [17], (2) estimate the rate of diffusion attributed to different theoretical aspects of the diffusion processes, b_0, external influence or innovativeness, and b_1, internal influence or interpersonal persuasion [15,16,18]. This model can be used to estimate rate of disease spread from a central source such as contaminated food or from infectious spread through interpersonal contact. In the social realm, one can use the model to estimate rate of adoption from a mass media advertisement or from interpersonal influence. Rate parameter estimates from both models for two diffusion datasets are provided in Table 1. Interpretation of these estimates is highly dependent on the time scale used to measure diffusion.

These rate parameter estimates can be used as outcomes to study factors associated with diffusion at the macro-level by comparing rates between groups and/or populations. For example, parameter estimates for different countries can be compared to study factors associated with the spread of behaviors in different countries. Modeling at this macro-level, however, is imprecise at best because it assumes perfect social mixing, everyone interacting with everyone else [19,20]. These macro models do not measure whether people who are connected to one another engage in the same behaviors. Geographers have devoted considerable attention to trying to determine whether innovations spread between contiguous areas.

Spatial Autocorrelation

Rather than just estimate rate of diffusion, spatial models measure whether artifacts, diseases, farming practices, and other behaviors spread between contiguous areas [2,21,22]. Proximity data are easy to obtain and are relatively unambiguous, thus providing a network of connections based on distance. Moran's I was an early model developed to test for spatial association, geographic clustering of adoption:

$$I = \frac{N \sum_{i}^{N} \sum_{j}^{N} D_{ij}(\boldsymbol{y}_i - \bar{y})(\boldsymbol{y}_j - \bar{y})}{S \sum_{i}^{N} (\boldsymbol{y}_i - \bar{y})^2} \,, \tag{3}$$

where N is the sample size, D a distance matrix (as proximities), y indicates adoption, and S the sum of the distances in the distance matrix. Moran's I measures the degree nodes that are connected to one another deviate from the average behavior in the network similarly or differently. Moran's I is high when connected nodes (positive elements of D) are either positively or negatively different from the average score. The statistical significance of Moran's I can be calculated in two ways: via permutation methods or analytically.

To use a permutation method to calculate the significance of Moran's I assume adoption (y_i) is randomly distributed and calculate I repeatedly to get a sample of estimates based on D and the number of adopters. If Moran's I calculated is significantly different than the random sample generated, Moran's I is considered significant (z-scores can be obtained). The logic then is to calculate the degree neighbors (however defined) have similar adoption behavior compared to that expected if adoption were distributed randomly. Variance estimators for Moran's I can be found in spatial statistics textbooks [21,23] and used to calculate exact significance tests. Moran's I is useful, and has been extended considerably, yet this approach often assumes that geographic proximity equates with communication and influence, which may not be true.

The spatial autocorrelation methodology was seen as a useful approach to measuring network autocorrelation, the bias inherent in a regression model when "y" appears as both the dependent and independent variable. Erbing and Young [24] wrote an influential paper on measuring network effects and using network autocorrelation methods. Dow [25,26] demonstrated the effects of network autocorrelation on estimate errors, and Doriean and others [27] found considerable bias in the point estimates and their standard errors. Exactly how network autocorrelation applied to diffusion of innovations was not clear since spatial autocorrelation measured diffusion at the macro level, but did not show whether specific individuals were more or less likely to adopt based on their network position. Further, spatial autocorrelation did not show how network structure influenced diffusion. To do so, we turn to network models.

Network Models

Figure 1 displays two networks from a study conducted in Cameroon among women in voluntary organizations [28]. Women were asked to name their friends in the organization in an attempt to determine if friendship ties were associated with contraceptive choices (they were). The diffusion network model posits that initial contraceptive choices would be made by some women based on their innovativeness and exposure to outside sources of influence such as their cosmopoliteness, media use, or greater need for the innovation. The new idea, and its practice, then spreads through the network as users persuade nonusers to adopt either by exhortation, entreaty, enticement, or example.

These influences are captured by an exposure or contagion model (Fig. 2), each individuals likelihood of adoption increases as the proportion (or number) of users in his/her personal network increases. Personal network exposure is the proportion or number of adopters in each person's network that provide information and influence with regard to some behavior. The equation for non-random mixing, or personal network exposure is:

$$E_i = \frac{\sum w_{ij} \boldsymbol{y}_j}{\sum w_i} , \qquad (4)$$

where w is the social network weight matrix, and y is vector of adoptions. For an individual who reported five contacts, network exposure (E_i) is the proportion of those contacts that have adopted (Fig. 2). When network exposure is measured on direct contacts it captures social influence conveyed through overt transmission of information, persuasion or direct pressure. Alternatively, exposure can be calculated by transforming the social network, W, to reflect other social influence processes. For example, W can be transformed to represent the degree of structural equivalence (similarity in network position) among people in the network. Exposure calculated on this network captures social influence conveyed via comparison to equivalent others such as by social comparison or competition [14]. Exposure can also be weighted by network properties such as centrality to reflect social influence by opinion leaders.

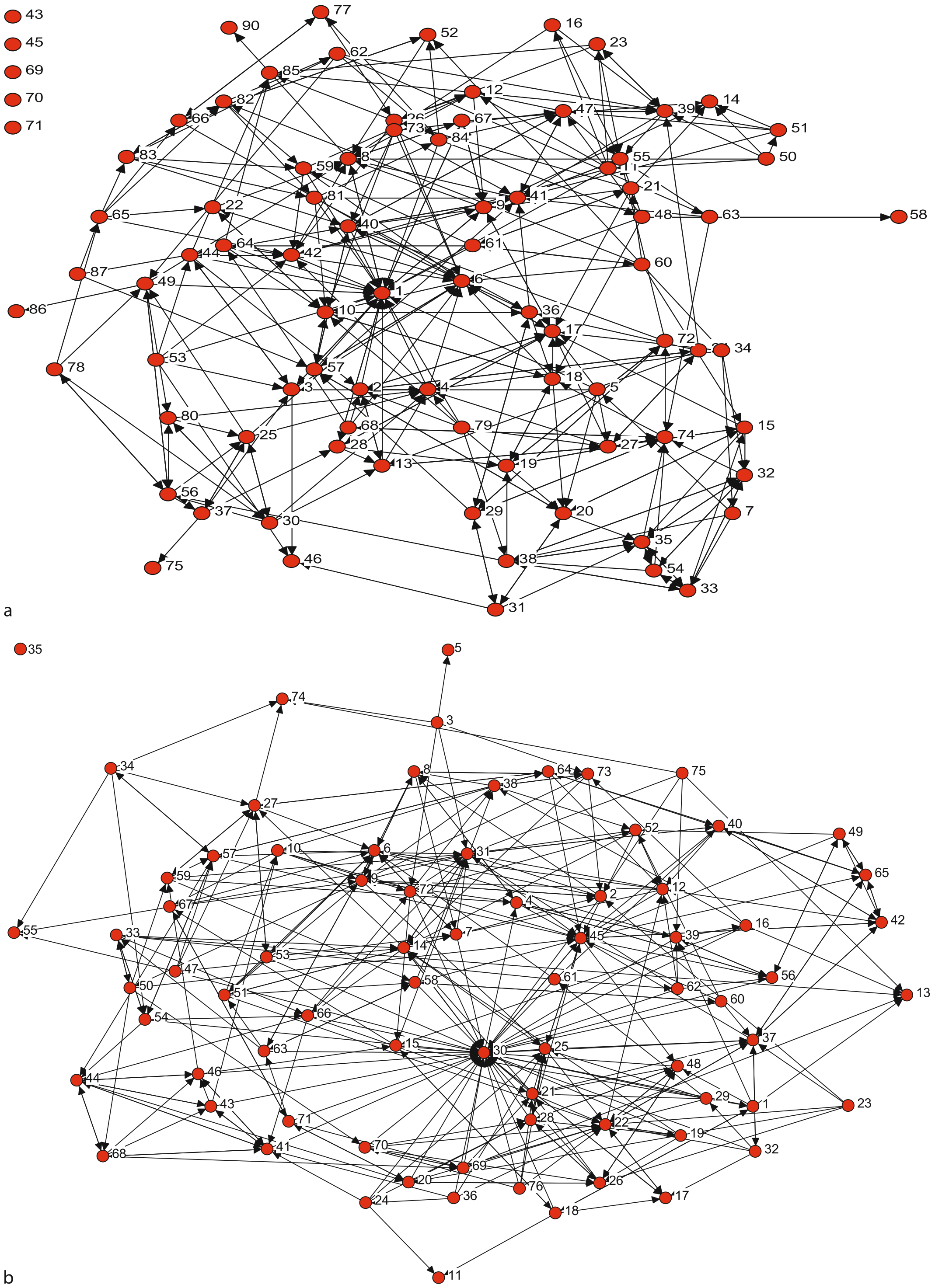
Social Networks, Diffusion Processes in, Table 2
Social network influence weightings

Relational	Positional	Central
1. Direct Ties	1. Percent Positive Matches (Tie Overlap)	1. Degree
2. Indirect Ties	2. Euclidean Distance	2. Closeness
3. Joint Participation in Groups or Events	3. Regular Equivalence	3. Betweenness
		4. Flow
		5. Integration/radiality
		6. Information
		7. Power

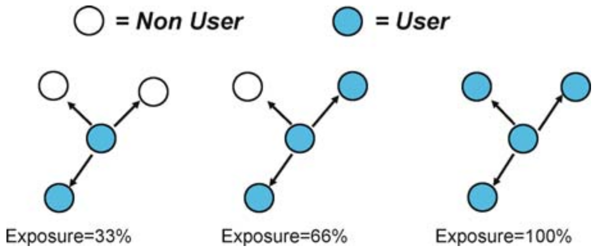
These three social influence processes are modeled with three different classes of network weight matrices (relational, positional, and central), constructed from the same social network data (Table 2). All three can be justified theoretically as sources of influence on adoption behavior and all three can be calculated various ways (there are at least 10 centrality measures). It is possible that all three operate for different people or at different times during the diffusion process.

In addition to the social influence process, a second dimension to these influence mechanisms is the weights attached to each based on social distance. For example, in relational influence models, different weights can be assigned to direct ties, ties-of-ties, and even the ties-of-ties-of-ties; in positional equivalence models, different weights can be assigned to those that are more equivalent than others [7]. A potential line of diffusion network research then is to compare different network weighting mechanisms to model and compare different social influence processes.

Diffusion was simulated through the two Cameroon networks in Fig. 1 to illustrate how network exposure and network structure influence diffusion. At each time period, adoption occurred for the non-adopter with the most nominations received, then network exposure was calculated, all nodes with exposure of 50% or higher were categorized as adopters and the process repeated. We compared diffusion in this network to that simulated in a net-



Social Networks, Diffusion Processes in, Figure 1
Networks 1 and 2 from the Cameroon Voluntary Association Study



Social Networks, Diffusion Processes in, Figure 2
Personal network exposure from direct contacts

work of the same size and density, but with links allocated randomly. Both conditions were averaged across 1000 runs. Figure 3 shows that, in Network 1, initially the diffusion trajectories are similar, but at about time 10, diffusion in the actual network accelerated.

The network accelerated diffusion because it is somewhat centralized (in-degree 21.7%) and once diffusion reaches the center of the network it can propagate rapidly. Notice that at about time 20 diffusion slowed, accelerated again at about 25, and then slowed from about time 30 to 40. These "fits and starts" are a product of the network structure: diffusion reaches pockets of inter-connectivity and spreads rapidly within these dense pockets, but slows between groups. Network 2 (Fig. 3) had more rapid and sustained diffusion because it was even more centralized (in degree, 47.2%). Note that in the spatial autocorrelation model adoptions were randomized to measure statistical significance and in this simulation, the network structure was randomized to illustrate its influence on the rate of diffusion.

Simulation assumptions regarding influences on adoption could easily be changed to achieve different outcomes. For example, when adoptions were assigned randomly, diffusion was constant in network one (and saturation lower) and similar to the random network in network two. The validity of these diffusion models rests partly on determining whether network exposure influences adoption. To that end, a number of empirical studies have been conducted to measure the degree social network exposure is associated with adoption.

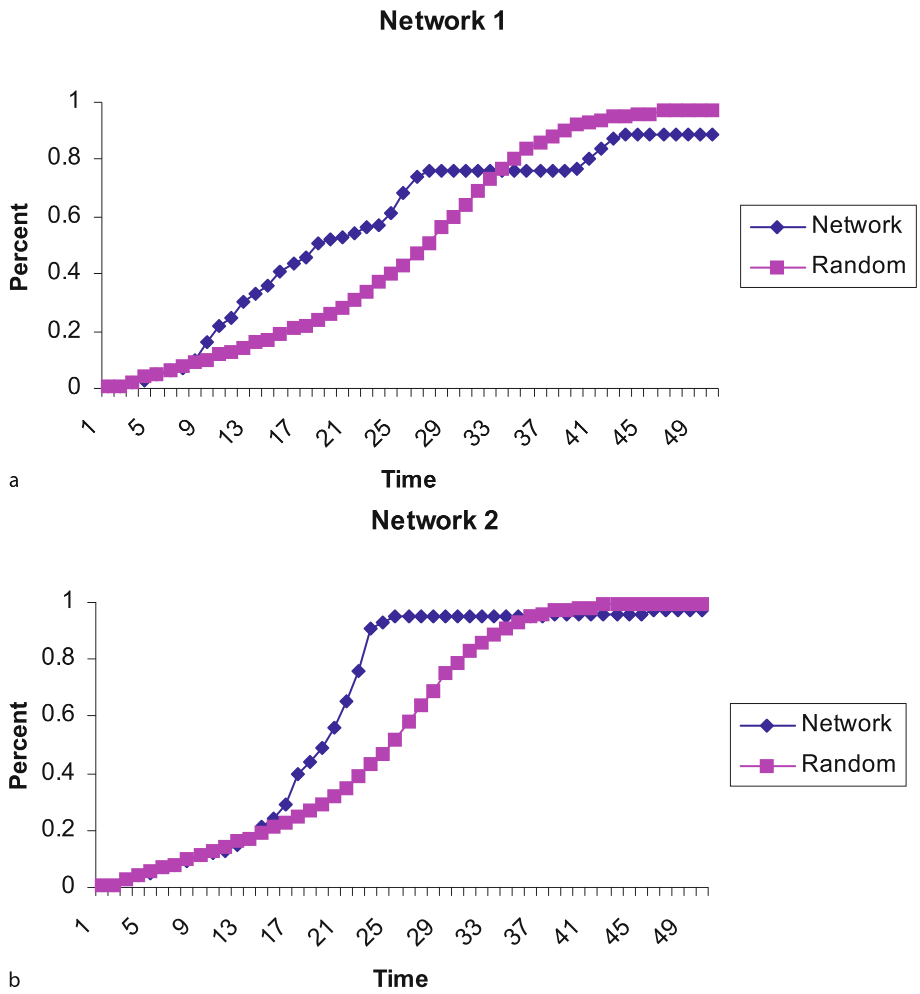
Empirical Studies

Empirical support for an association between individual own behavior and that of his/her peers can be found throughout the behavioral sciences literature. While many scholars assume adoption is associated with network exposure, few studies have traced an innovation through a network of social contacts to empirically validate this proposition. The lack of data on diffusion within an entire network stems largely from the difficulty of trying to collect data over a time period long enough for diffusion to occur. Consequently, most studies have relied on retrospective data which introduces some but not much bias [29,30]. It has also meant that several scholars have re-analyzed two studies that collected network and adoption data, (1) medical innovation study [12], reanalyzed by Burt [14], Marsden and Podolny [31], Strang and Tuma [32], Valente [7,33], and Van den Bulte and Lillien [34]; (2) Korean family planning study [35], reanalyzed by Dozier [36], Valente [7,33], Montgomery and Chung [37], and Kohler [38]. Recent studies in the fields of reproductive health [39] and substance abuse [40] have provided new data, but these classics remain classic.

Because collecting complete network data can be difficult, most empirical research has been ego-centric [41,42], based on respondent reports of their behavior and that of their network peers who are not necessarily connected to one another and not interviewed. Social influence is often based on respondent reports of perceptions of peer behavior or perceptions of peer influence [43,44]. Comparison of exposure scores based on respondent perceptions with alters' reports in one study found that perceptions were more strongly related to behavior than exposure based on alter reports [28,37,45].

Sociometric studies interview members of a bounded community and attempt to gather information from everyone in the community (typically conducted in schools, organizations, and small communities) and record their time of adoption [1,12,46,47,48]. Sociometric studies are useful for understanding how an innovation flows within the community and how certain network structural variables influence the diffusion process. Sociometric data capture network influences by the alters' reports since they were also interviewed. For example, sociometric studies can determine whether structural positions such as centrality are associated with adoption and/or whether centralization is associated with more rapid diffusion [7].

A number of recent diffusion network studies have been cross-sectional and in many cases retrospective involving only one time point. For example, a study in Thailand by Entwisle and others [49] found that contraceptive choices made by early adopters contributed significantly to the contraceptive choices made by later adopters (also see [35]). Valente and others [28] collected sociometric data on contraceptive use among women in voluntary associations in Cameroon and showed that perceptions of these friends' behavior, and in particular perceptions that these friends encouraged contraceptive use, were significantly associated with behavior. In general, statistical ana-



Social Networks, Diffusion Processes in, Figure 3
Simulated diffusion in two networks, each compared to random networks of the same size and density. Network 2 is more centralized than Network 1

lyzes use the following model:

$$\log \frac{\Pr(\boldsymbol{y}_t = 1)}{(1 - \Pr(\boldsymbol{y}_t = 1))} = \alpha + \sum B_k X_k + B_{(k+1)} \omega \boldsymbol{y}_t , \quad (5)$$

where $\boldsymbol{y}$ is a binary vector of adoption behavior, α is the intercept, β_k are parameter estimates for vectors of K socio-demographic characteristics (Xs), and ω represents the social network matrix. The $\omega \boldsymbol{y}_t$ term represents the calculation of contemporaneous network exposure and this vector is usually divided by a count of the number of nominations sent (alternatively the number of nominations can be entered into the regression separately).

Significant estimates for β_{k+1} indicate contagion effects by showing that network exposure is associated with adoption. The variances for these estimates, however, are usually biased since the observations are not independent and hence the errors in prediction are not independent. One partial solution is to obtain robust estimates by controlling for clustering. Clustering is the degree that elements from the same cluster are similar compared to those of different clusters. For example, two individuals chosen at random from the same organization are more likely to be similar than two chosen at random from different organizations. Table 3 reports regression results of the Cameroon data with and without correction for clustering. Without correction, network exposure is strongly and significantly associated with adoption, but with the correction it is only marginally statistically significant (p=.04). Controlling for clustering is particularly important in network exposure models because network choices are often restricted to the cluster.

Social Networks, Diffusion Processes in, Table 3
Logistic regression on the likelihood of contraceptive behavior on controls and network exposure with and without correction for clustering (N=555; Groups=9)

	Contraceptive Method Use			
	Without Correction		With Correction	
	Adjusted		Adjusted	
	Odds Raios	P-value	Odds Raios	P-value
Age	0.97	0.001	0.97	0.010
Education	0.91	0.247	0.91	0.184
Possessions	1.39	0.000	1.39	0.000
Network Exposure	1.14	0.005	1.14	0.047

Even with clustering controlled, social influence as measured through social networks seems to be strongly associated with behavior. For example, a school-based sociometric study was conducted by Alexander and colleagues [50] using the Adolescent health data [51] to show that students with a majority of network ties who were smokers were almost two times as likely to smoke themselves with an additional two times greater likelihood of smoking for those with best friends who smoke. Clustering was controlled and the multi-level model accurately captured micro-level effects within the context of macro-level influences. The study measured the influence of peers on smoking while conditioning on the smoking rate within the school [50].

Estimating the network exposure (autocorrelation) term with a multi-level model can provide contagion estimates across settings and estimate the degree it varies between settings (i. e., communities, schools, organizations, and so on). The models are incomplete however, because there may be factors that influence both adoption and choice of social network contacts. For example, the decision to smoke and to nominate friends who smoke may both be a function of delinquency or rebellion. Hence an association between behavior and peer behavior can be spurious. Testing social influence with network methods then requires longitudinal data involving at least two time points. Boulay and Valente [52] collected data among women in three villages of Nepal and found that having discussion partners who used contraception influenced information seeking behavior and contraceptive choice. Having data from two time points allows testing of a simple dynamic model on adoption:

$$\log \frac{\Pr(\boldsymbol{y}_t = 1)}{(1 - \Pr(\boldsymbol{y}_t = 1))} = \alpha + \sum B_k X_k + B_{(k+1)} \omega_t \boldsymbol{y}_t + B_{(k+2)} \omega_{(t-1)} \boldsymbol{y}_{(t-1)}, \quad (6)$$

where y is a binary indicator of behavior, α is the intercept, β_k are parameter estimates for vectors of K socio-demographic characteristics (Xs), and ω represents the social network matrix. A positive and significant β_{k+2} indicates that respondents with high network exposure at baseline were more likely to adopt at time two. A positive and significant β_{k+1} indicates that change in network exposure is associated with change in behavior. This may indicate contagion but still may be a product of some omitted factor. Panel data collected at two time periods is adequate for most research needs, and can provide evidence of network influences on behavior. However, since there is often a considerable time between the two measures, many factors may account for simultaneous change in behavior and network exposure. To cope with this threat, data can be collected on time of adoption, expanding the micro level dynamic analysis by using event history analysis [53].

Event History Analysis

Event history analysis techniques have been developed to analyze data with a substantive number of time points, estimating coefficients with maximum likelihood estimators [32,53,54,55,56]. There are two types of event history analysis, discrete time, in which the outcome is binary, and continuous time, in which the outcome is time-to-an-event. Since diffusion occurs over time, there is an explicit time dimension in diffusion studies captured by both discrete and continuous time models. The time of adoption variable is the dependent variable and may be influenced by both time-varying and time-constant factors. Some individuals may not have adopted by the time of data collection giving rise to time-censored observations. Right censoring occurs when data are collected before the innovation has finished diffusing or does not diffuse to all members of the community or study. Left censoring occurs when the data are incomplete at the beginning of the process. For example, adoption data for the period 1993 to 2000 may have some people who adopted in 1989–1992 classified as 1992 adopters.

There are a variety of event history techniques including hazard models developed in epidemiology used to understand the hazard or risk to disease or injury over time. Hazard and/or event history analysis generally requires that the data be reshaped from simple observations to a case-time format such that there is a case in the data for each individual at each time period of study up to and including that person's time of adoption. The time-varying and time-constant independent variables are included in each case as well as a binary indicator for whether the individual adopted or not (got sick or not).

Maximum-likelihood estimation can determine whether the independent variables are associated with the dependent variable (adopt/not adopt) [57]. A study of 100 people with an average adoption time of seven translates into 700 person-time cases. Each person-time case has a variable for the network exposure at that time period plus an indicator for whether the person adopted or not (plus additional time-constant and time-varying covariates as desired). The event history model is:

$$\log \frac{\Pr(\boldsymbol{y}_t = 1)}{(1 - \Pr(\boldsymbol{y}_t = 1))} = \alpha + \sum B_j X_j + \sum B_{kt} X_{kt} + \sum B_{(k+1)} \omega \boldsymbol{y}_t \,, \tag{7}$$

where y is a binary indicator of behavior, α is the intercept, β_j are parameter estimates for vectors of J socio-demographic characteristics (X_j), and β_{kt} are parameter estimates for the matrix of time-varying socio-demographic characteristics (X_{kt}) and ω represents the social network weight matrix, and t a time indicator. Note here we have assumed a static (constant) network. Standard statistical packages allow testing of event history or survival data in a relatively straightforward manner, once the data are reformatted. Event history analysis requires the construction of exposure matrices for each time period which can be a formidable task particularly if one uses more than one network weight matrix.

Marsden and Podolny [31] used event history analysis and tested network exposure's association with adoption in the medical innovation data. Results showed that exposure was not associated with adoption in that study. Strang and Tuma [32] revisited the issue with the same data by postulating time variance in network influence, (i. e., how much lag time, if any, is there in the influence). They found evidence of contagion. Van den Bulte and Lillien [34] supplemented the medical innovation data with archival data on media promotion by pharmaceutical firms at the time of the original study and showed that network contagion effects disappear once these data are added. Their analysis demonstrates the importance of omitted variables when studying diffusion through networks. The rapid diffusion measured in the medical innovation study indicates that contagion was probably not the primary factor driving diffusion.

Event history analysis of the three classic diffusion network datasets has been conducted [58]. The analysis controlled for within village and within person covariation, and terms for time and a logistic transformation of time were included to control for macro-level effects. Terms for infection and susceptibility [32,59] were included to measure whether adoption by central individuals (high in-degree) influenced subsequent adoption, infection, and whether centrality (out-degree) influenced a person's likelihood to adopt as diffusion occurred, susceptibility. In-degree and out-degree were also included in the model. Two network exposure terms were computed, direct ties and structural equivalence. Structural equivalence was computed as in Burt's [14] measure, Euclidian distance raised to the 16th power. For this analysis, network exposure was calculated using contemporaneous measures since two of the datasets recorded adoption in one-year intervals. Two control variables representing individual characteristics were included. Analysis was conducted only on those who adopted. The following model was estimated:

$$\log(\Pr(\boldsymbol{y}_t = 1)) = \alpha + \sum \beta_{lm} X_{lm} + \sum \beta_{lmt} V_{lmt} + \sum \beta_{lmt} \omega_s \boldsymbol{y}_t + \lambda_{lm1} C_D(\boldsymbol{y}_+) + \lambda_{lm2} C_D(\boldsymbol{y}_+) \,, \tag{8}$$

where y is a binary indicator of behavior, α is the intercept, Xs are vectors or time-constant socio-demographic and network characteristics, V represents vectors of time-varying terms, in this case time and its transformation, ω_s represents the social network matrices, and λ estimates the effects of centrality degree variables multiplied by the time varying proportion of adopters in the network (infection and susceptibility). Results were mixed, but seem to indicate that both infection and susceptibility effects are present. In all three datasets, infection is positively associated with adoption indicating that as those with high in-degree adopt, the likelihood others in the network will adopt increases. In two studies, Brazilian farmers and Korean women, susceptibility is associated with adoption indicating that those with a high number of nominations sent are more likely to adopt as the innovation diffuses. Ties sent and received are marginally associated with adoption, and only for the Brazilian data is exposure, through structural equivalence, associated with adoption. These results, however, change dramatically when non-adopters are included or when a term for the average exposure at each time period is included such that infection and susceptibility effects disappear.

The event history analysis approached has also been used by Montgomery and others (2001) using ego-centric data to study network exposure's influence on contraceptive use in Ghana. Current analysis of four rounds of data over two years has shown that contraceptive use is strongly associated with use by social network peers. The Ghana field study provides some of the most conclusive evidence of the magnitude of social influence on behavior change by showing that as the number of social network contacts who use contraceptives increases, the likelihood of contraceptives use by ego also increases. Of all the variables, the network exposure variables were the most significant influences on contraceptive adoption. Another longitudinal field study in Kenya found similar results, again based on ego-centric network data [60].

Montgomery and others [61] also report preliminary analysis in which network influences are weighted by tie characteristics such as the frequency of communication. They found that adding these weights did not change the strength of peer influence. Similar results have been reported in Valente [7] and Valente and Saba, p. 109 in [43]. Consequently, it seems that the influence of social networks on behavior (contraceptive use in these cases) seems broad in nature, and are not conditioned on specific factors such as the frequency of communication between dyads or the their socio-demographic equality. These factors may play a strong, and even pervasive role in determining who is connected to whom [62], but they do not seem to determine the degree of influence social contacts provide.

Network exposure and adoption may not always be strongly correlated for a number of reasons. First, exposure may not be associated with adoption for everyone, but may be most influential during the middle stages of diffusion, when awareness is high, but uncertainty about its relative advantages is also high [63]. Exposure may have less of an effect early in the process when there are few adopters and obvious advantages to waiting; and late in the process when most people have a majority of adopters in their personal network anyway. Second, individuals may have varying thresholds to adoption such that some are innovative and others are not [19]. Valente [7,33] posited a social network threshold model in which contagion (majority rule) is a special case. Most simulation models assume majority influence on adoption decisions as was done in the beginning of this chapter. It is reasonable, however, to expect that individuals vary in the amount of network exposure needed to adopt an innovation. Disproving thresholds may not be possible, but construct validity for the concept has been demonstrated [33]. Valente and Saba [43] replicated the threshold model using ego-centric data and showed that people with a minority of network members using contraception had higher campaign recall indicating that the media campaign could substitute for interpersonal sources of influences. If thresholds vary, network exposure is needed for people to reach those thresholds, if they do not and the special case of contagion exists, network exposure will determine when individuals adopt.

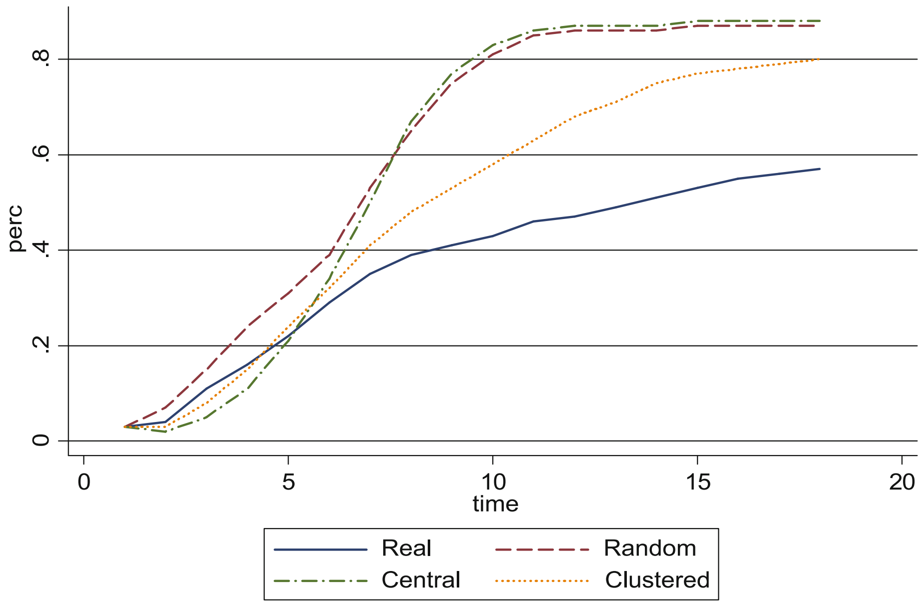
In spite of the impressive list of studies showing some support for an association between individual behavior and network exposure, and the theoretical simulations of network structure and thresholds, significant work remains to be done. Most scholars and lay people would agree that social networks influence behavior. The barriers to demonstrating this effect, however, have been challenges of data collection and agreement on appropriate statistical methodology. The most commonly analyzed dataset, Medical Innovation, is 45 years old, consists only of 125 respondents and arguably is not a diffusion study at all. The limitation of available data has forced researchers to rely on simulations and agent based modeling approaches to understand theoretical mechanisms driving diffusion.

Agent Based Models

Agent based models (ABM) are computer simulations created to understand how actors (usually though not necessarily people) behave given a set of preexisting conditions and rules for their behavior [64]. A researcher can hypothetically pose a population of people who have some degree of connectedness and some structure to those connections. Model parameters may be set that vary those structural conditions or they may be taken as fixed and other properties varied. To model the diffusion of innovations one might be interested in how network structure affects the rate of diffusion. So a simulation can be created which varies network structure and then simulates the spread of an innovation within those structures. One could also vary other properties of the diffusion by varying the number of initial adopters or the transmissibility of the innovation.

To illustrate, an empirical dataset was used to simulate theoretical network structures and potential diffusion within those structures. Data were collected among adolescents in ninth grade and they were asked to write the names of their five closest friends. A computer program was written to match the names to the roster in the school so a network could be constructed from the data. The network had 150 nodes with 327 friendship links. The agent based model had 2 parameters: (1) the initial 7% seed adopters would vary under 3 conditions: (a) randomly selected seeds, (b) the most central nodes, and (c) peripheral nodes; (2) the type of network structure would vary by 4 conditions (a) the real network, (b) a random one with the same links, (c) a centralized one, and (d) a clustered one.

A simple diffusion model was then run setting a low threshold for adoption such that at each time interval after the initial seeding those nodes connected to 15% other



Social Networks, Diffusion Processes in, Figure 4
Comparison of simulated diffusion trajectories for a real network, and 3 simulated ones based on the real network's density. The three simulated networks are random links, centralized links (centralization = 3%), and clustered links (clustering coefficient=12%). Diffusion occurs most rapidly in random and centralized networks

nodes would become adopters. The model was run for 20 time periods which was generally long enough for diffusion to occur in most situations. After a run of 20 time periods was completed, results were tabulated and then the simulation for each condition run 25 times and the results aggregated. The model was run 25 times for each condition to ensure that the results were not the product of some anomaly. It is customary in agent based models to run the simulations for hundreds or thousands to times to guarantee the results are robust.

Figure 4 shows illustrative findings by plotting the diffusion trajectories in the different conditions. The data show that simulated diffusion occurs most rapidly in the random and centralized networks. It is slowest in the real network. The network structure of real networks retards diffusion because of homophily – the tendency for people to be friends with others like themselves. This homophily tendency creates clustered networks so that diffusion gets trapped in these pockets of interconnectivity. Real networks also retard diffusion because they are symmetric (If $A \rightarrow B$ then $B \rightarrow A$) and so adoption by a pair leads to reinforcement, not more diffusion. Bridges are needed to carry the innovation across the gap between clusters and these bridges act as bottlenecks that slow diffusion.

Statistical analysis can be conducted on the data to determine the effect on diffusion. Outcomes can be the final cumulative percent of adopters (the prevalence), or the number of time periods till diffusion has reached 50%, or the rate of adoption estimated using one of the mathematical models described above. In these data we find that all 3 simulated networks have more rapid diffusion than the real network, and we find that seeding the network with central adopters accelerates diffusion while seeding it with peripheral actors slows diffusion.

Conclusions

Much progress has been made since 1943 when Ryan and Gross first laid the foundation for diffusion of innovations theory. Rogers [1] chronicled the many studies conducted since then and helped shaped a general diffusion model with wide applicability now being renewed and reinvigorated with fresh theory and analytic models. Overall, results indicate that social network influences on behavior are important and have consequences for the health and well being of populations and individuals. These new insights have shed light on important aspects of how new ideas and practices spread within and between communities.

Along with new insights have come new questions and new perspectives to be addressed. It is clear that a lack of data on time of adoption coupled with information on network relations has hampered developments. Few diffusion or behavioral studies collect information on networks, conversely few network studies record time of adoption. There are advantages to marrying these two ideas, however, and future research will hopefully try to collect both types of data.

It is also clear that our understanding of how diffusion occurs is still somewhat limited. The Medical Innovation data have often been used to demonstrate the importance of networks in adoption yet analyses by Valente [7] and by Van den Bulte and Lillien [34] have shown that contagion via social influence in this setting was unlikely. Given the number of confounding factors, and some of the data requirements, it may be prohibitively difficult to substantiate the role of social networks in innovation adoption via survey methods alone. Purposively intervening on social networks, however, may prove to be a fruitful avenue of research. If network-based interventions can be used to accelerate innovation diffusion, then a stronger case can be made for the importance of social contagion in the diffusion process.

Nonetheless, it is clear that networks are important influences on behavior since most people acknowledge that they receive information and influence via their social networks and that they model the behavior of others. What is less clear is how to capture that influence in quantitative terms that mimic the theoretical progress made in the network field. Further, verbal accounts on how people make decisions and adopt behaviors usually reveal non-linearities, chance circumstances, and whims that are not independent of networks, but not easily captured in social influence models.

The link between micro and macro levels of analysis represents an opportunity for study of diffusion processes. The opportunity lies in the fact that multi-level modeling techniques enable the separation of micro-level network exposure influences from macro-level contextual factors. Yet both are social network influences and both represent elements of the diffusion paradigm. It is hoped that by controlling for contextual effects we don't "throw the baby out with the bathwater," by eliminating the micro-level influences that provide expressions for those contextual effects.

In spite of controls for macro-level contextual effects, micro-level associations between peer network behavior, and those of respondents are still sometimes strong. Debate remains about the meaning of these associations, is it peer influence, peer selection, or further contextual effects? More rigorous studies may eventually tease this out, in the interim, better study designs and interventions will need to be created. This review has attempted to point out some of the challenges diffusion scholars face and some of the promising new directions it may take.

Scholars have turned to agent based modeling as a technique to investigate diffusion properties. Agent based models (ABM) allow the theoretical exploration of how network structure can be varied and shown to influence diffusion. ABM also enables us to study diffusion parameters such as thresholds, transmissibility, or the importance of initial adopters on diffusion speed and prevalence. Until empirical data can be gathered and analyzed ABM represent a promising approach to understanding how diffusion processes occur on networks.

Future Directions

Future developments in diffusion processes are likely to occur along two interrelated fronts. First, increasingly scholars are using network data to propose and implement social change practices [65]. These activities may involve organizational change in the business context or health promotion and treatment in the medical context. These network based interventions provide the opportunity to change networks and measure the resulting changes in performance or behavior. Thus, scholars can more definitively link changes in networks to changes in outcomes.

Second, considerable progress has been made in advancing statistical models that measure network evolution and behavioral changes that account for multiple structural influences [66]. For example, scholars can test the degree of similarity in behaviors while simultaneously estimating the influence of social network characteristics such as whether two people nominate one another or are part of a linked triad. These models (known as P* or exponential random graph models) provide a statistical estimation of concepts such as network influence, and network selection.

Acknowledgments

Support for this research was provided by NIDA grant #DA10172. Some portions of this chapter appeared in [58].

Bibliography

1. Rogers EM (2003) Diffusion of innovation, 5th edn. The Free Press, New York
2. Hägerstrand T (1967) Innovation diffusion as a spatial process. University of Chicago Press, Chicago
3. Brown L (1981) Innovation diffusion: a new perspective. Methuen, New York
4. Robertson TS (1971) Innovative behavior and communication. Holt, New York

5. Bailey NTJ (1975) The mathematical theory of infectious diseases and its applications. Charles Griffen, London
6. Morris M (1993) Epidemiology and social networks: modeling structured diffusion. Sociol Methods Res 22:99–126
7. Valente T (1995) Network models of the diffusion of innovations. Hampton Press, Cresskill
8. Bael GM, Bohlen JM (1955) How farm people accept new ideas. Cooperative Extension Service Report 15, Ames
9. Katz E, Levine ML, Hamilton H (1963) Traditions of research on the diffusion of innovation. Am Sociol Rev 28:237–253
10. Ryan R, Gross N (1943) The diffusion of hybrid seed corn in two Iowa communities. Rural Sociol 8(1):15–24
11. Valente TW, Rogers EM (1995) The origins and development of the diffusion of innovations paradigm as an example of scientific growth. Sci Commun 16(3):242–73
12. Coleman JS, Katz E, Menzel H (1966) Medical innovation: a diffusion study. Bobbs Merrill, New York
13. Rogers EM (1957) Personality correlates of the adoption of technological practices. Rural Sociol 22:267–268
14. Burt R (1987) Social contagion and innovation: cohesion versus structural equivalence. Am J Sociol 92:1287–1335
15. Bass FM (1969) New product growth for model consumer durables. Manag Sci Ser a-Theory 15(5):215–227
16. Hamblin RL, Jacobsen RB, Miller JLL (1973) A mathematical theory of social change. Wiley, New York
17. Mahajan V, Peterson RA (1985) Models of innovation diffusion. Sage, Newburg Park
18. Valente T (1993) Diffusion of innovations and policy decision-making. J Commun 43(1):30–41
19. Granovetter M (1978) Threshold models of collective behavior. Am J Sociol 83:1420–1443
20. Van den Bulte C, Lillien GL (1997) Bias and systematic change in the parameter estimates of macro-level diffusion models. Marketing Sci 16:338–353
21. Cliff A, Ord JK (1981) Spatial processes: models and applications. Pion, London
22. Griffith DA et al (1999) A casebook for spatial statistical data analysis: a compilation of analyses of different thematic data sets. Oxford University Press, New York
23. Bailey TC, Gatrell AC (1995) Interactive spatial data analysis. Longman, Essex
24. Erbing L, Young A (1979) Individuals and social structure: contextual effects as endogenous feedback. Sociol Methods Res 7:396–430
25. Dow M, Burton ML, White DR (1982) Network autocorrelation: a simulation study of a foundational problem in regression and survey research. Soc Netw 4:169–200
26. Dow M (1986) Model selection procedures for network autocorrelated disturbance models. Sociol Methods Res 14:403–422
27. Doreian P, Teuter K, Wang C (1984) Network autocorrelation models: some Monte Carlo results. Sociol Methods Res 13:155–200
28. Valente TW et al (1997) Social network associations with contraceptive use among Cameroonian women in voluntary associations. Soc Sci Med 45:677–687
29. Coughenour CM (1965) The problem of reliability of adoption data in survey research. Rural Sociol 30:184–203
30. Nischan P et al (1993) Comparison of recalled and validated oral contraceptive histories. 138(9):697–703. Am J Epidemiol 189(9):697–703
31. Marsden PV, Podolny J (1990) Dynamic analysis of network diffusion processes. In: Weesie J, Flap H (eds) Social networks through time. ISOR, Utrecht
32. Strang D, Tuma NB (1993) Spatial and temporal heterogeneity in diffusion. Am J Sociol 99:614–639
33. Valente T (1996) Social network thresholds in the diffusion of innovations. Soc Netw 18:69–89
34. Van den Bulte C, Lillien GL (2001) Medical innovation revisited: social contagion versus marketing effort. Am J Sociol 106:1409–1435
35. Rogers EM, Kincaid DL (1981) Communication Networks: A new paradigm for research. Free Press, New York
36. Dozier DM (1977) Communication networks and the role of thresholds in the adoption of innovations. Ph D Thesis, Stanford University
37. Montgomery MR, Chung W (1999) Social networks and the diffusion of fertility control in the Republic of Korea. In: Leete R (ed) Dynamics of values in fertility change. Oxford University Press, Oxford
38. Kohler HP (1997) Learning in social networks and contraceptive choice. Demography 34:369–383
39. Casterline J (2001) Diffusion processes and fertility transition: selected perspectives. National Academy Press, Washington
40. Neiagus A et al (2001) HIV risk networks and HIV transmission among injecting drug users. Eval Program Plan 24:221–226
41. Marsden PV (1987) Core discussion networks of Americans. Annu Rev Sociol 52:122–131
42. Marsden PV (1990) Network data and measurement. Annu Rev Sociol 16:435–463
43. Valente T, Saba W (1998) Mass media and interpersonal influence in a reproductive health communication campaign in Bolivia. Commun Res 25:96–124
44. Valente T, Vlahov D (2001) Selective risk taking among needle exchange participants in Baltimore: implications for supplemental interventions. Am J Public Health 91:406–411
45. Urberg KA, Degirmencioglu SM, Pilgrim C (1997) Close friend and group influence on adolescent cigarette smoking and alcohol use. Dev Psychol 33:834–844
46. Becker MH (1970) Sociometric location and innovativeness: reformulation and extension of the diffusion model. Am Sociol Rev 35:267–282
47. Scott J (2000) Network analysis: a handbook, 2nd edn. Sage, Newbury Park
48. Wasserman S, Faust K (1994) Social networks analysis: methods and applications. Cambridge University Press, Cambridge
49. Entwisle B et al (1996) Community and contraceptive choice in rural Thailand: a case study of Nang Rong. Demography 33:1–11
50. Alexander C et al (2001) Peers, schools, and adolescent cigarette smoking. J Adolesc Health 29(1):22–30
51. Bearman PS, Jones J, Udry JR (2000) National longitudinal study of adolescent health: research design. http://www.cpc.unc.edu/projects/addhealth/data
52. Boulay M, Valente TW (2005) Dynamic sources of information and dissonance in the discussion networks of women in rural Nepal. J Health Commun 10:519–536
53. Tuma NB, Hannan MT (1984) Social dynamics: models and methods. Academic Press, New York
54. Allison PD (1984) Event history analysis. Sage, Newberry Park
55. Bartholomew DJ (1982) Stochastic models for social processes. Wiley, New York

56. Teachman JD, Hayward MD (1993) Interpreting hazard rate models. Sociol Methods Res 21(3):340–371
57. Eliason SR (1979) Maximum likelihood estimation: logic and practice. Sage, Newburg Park
58. Valente T (2005) Models and methods for innovation diffusion. In: Carrington PJ, Scott J, Wasserman S (eds) Models and methods in social network analysis. Cambridge University Press, Cambridge
59. Myers DJ (2000) The diffusion of collective violence: infectiousness, susceptibility, and mass media networks. Am J Sociol 106:173–208
60. Kohler H-P, Behrman JR, Watkins SC (2001) The density of social networks and fertility decisions: evidence from South Nyanza District, Kenya. Demography 38:43–58
61. Montgomery MR et al (1999) Social networks and contraceptive dynamics in Southern Ghana. Paper presented at the annual meeting of the Population Association of America. Washington DC
62. White K, Watkins SC (2000) Accuracy, stability and reciprocity in informal conversational networks in rural Kenya. Soc Netw 22:337–356
63. Carley KM (2001) Learning and using new ideas: a socio-cognitive approach. In: Casterline J (ed) Diffusion processes and fertility transition: selected perspectives. National Academy Press, Washington
64. Epstein JM (2006) Generative social science. Princeton University Press, Princeton
65. Valente TW, Fosados R (2006) Diffusion of innovations and network segmentation: the part played by people in the promotion of health. J Sex Transm Dis 33:S23-S31
66. Robins G, Pattison P, Kalish Y, Lusher D (2007) An introduction to exponential random graph (p*) models for social networks. Soc Netw 29:173–191

Social Networks, Exponential Random Graph (p^*) Models for

GARRY ROBINS
School of Behavioural Science, University of Melbourne,
Melbourne, Australia

Article Outline

Exponential random graph models, also known as p^* models, constitute a family of statistical models for social networks. The importance of this modeling framework lies in its capacity to represent social structural effects commonly observed in many human social networks, including general degree-based effects as well as reciprocity and transitivity, and at the node-level, homophily and attribute-based activity and popularity effects. The models can be derived from explicit hypotheses about dependencies among network ties. They are parametrized in terms of the prevalence of small subgraphs (configurations) in the network and can be interpreted as describing the combinations of local social processes from which a given network emerges. The models are estimable from data and readily simulated. Versions of the models have been proposed for univariate and multivariate networks, valued networks, bipartite graphs and for longitudinal network data. Nodal attribute data can be incorporated in social selection models, and through an analogous framework for social influence models.

The modeling approach was first proposed in the statistical literature in the mid-1980s, building on previous work in the spatial statistics and statistical mechanics literature. In the 1990s, the models were picked up and extended by the social networks research community. In this century, with the development of effective estimation and simulation procedures, there has been a growing understanding of certain inadequacies in the original form of the models. Recently developed specifications for these models have shown a substantial improvement in fitting real social network data, to the point where for many network data sets a large number of graph features can be successfully reproduced by the fitted models.

Glossary

Alternating independent-2-paths A parameter (and statistic) in new specification models; a particular combination of *k-independent-2-path* counts into the one statistic.

Alternating k-stars A Markov parameter (and statistic) in the new specification models; a particular combination of Markov k-star counts into the one statistic; equivalent to *geometrically weighted degree counts*; useful for modeling the degree distribution.

Alternating k-triangles A parameter (and statistic) in the new specification models; a particular combination of *k-triangle* counts into the one statistic; equivalent to *weighted shared partners*.

Cyclic triad A Markov graph configuration: in a directed network, ties *ij*, *jk* and *ki* are observed among actors i, j, and k.

Degeneracy (or near-degeneracy) When a model implies that very few distinct graphs are probable, often only empty or complete graphs; degenerate models cannot be good models for social network data.

Dependence assumption Theoretical assumption about dependencies among possible network ties; determines the type of parameters in the model.

Dyad independence Assumes that dyads are independent of one another; the model includes edge and reciprocity parameters, and possibly also node or dyad attributes.

Dyad-wise shared partners A parameter (and statistic) in the higher order models; equivalent to *alternating independent 2-paths*.

Edge-wise shared partner distribution Distribution of the number of dyads who are themselves related and who have a fixed number of shared partners.

Edge-wise shared partners A parameter (and statistic) in the higher order models; equivalent to *alternating k-triangles*.

Geometrically weighted degree counts A statistic (and parameter) in the new specification models: a sum of degree counts with geometrically decreasing weights; equivalent to *alternating k-stars*.

Homogeneity assumption Assumption about which parameters to equate, to make a model identifiable.

***k*-independent-2-paths** Configurations in the higher order models; equivalent to *k*-triangles but without the base.

***k*-in-star** A Markov graph configuration: in a directed graph, *k* arcs are directed to the one actor.

***k*-out-star** A Markov graph configuration: in a directed graph, *k* arcs are expressed by the one actor.

***k*-triangle** A configuration in higher order models; in a non-directed graph, the combination of *k* triangles, each sharing the one edge (the base of the *k*-triangle).

***k*-star** A Markov graph configuration: in a non-directed graph, *k* edges are expressed by the one actor.

Markov dependence assumption Introduced by Frank and Strauss [9], proposes that, conditional on the rest of the graph, two possible ties are independent of each other unless they share an actor.

Mixed-star A Markov graph configuration: a two path in a directed graph.

Monte Carlo Markov chain maximum likelihood estimation (MCMCMLE) Method of estimation based on computer simulation; more principled than pseudolikelihood.

Network configuration A small subgraph that may be observed in the data and that is represented by parameters in the model: e. g. reciprocated ties, triangles.

Parameters Relate to specific network configurations that may be observed in the graph; a large positive parameter is interpreted as the presence of more of the configurations than might be expected from chance (given the other effects in the model); a large negative parameter signifies the relative absence of the configuration.

Partial dependence assumption Assumption for dependencies among possible ties created by the presence of other ties; permits models with higher order configurations than Markov configurations.

Pseudo-likelihood estimation An approximate method of estimation using logistic regression; does not produce reliable standard errors.

p_1 Model An early dyad independence model, including popularity and expansiveness effects.

p_2 Model Elaboration of p_1 model, where popularity and expansiveness effects are random, and independent variables may be used to predict ties.

Simple random graphs, Bernoulli graphs, Erdös–Rényi graphs Assume that edges are independent of one another and are observed with a given probability.

Social circuit dependence Two possible ties are conditionally dependent when, if observed, they would create a 4-cycle.

Transitive triad A Markov graph configuration: in a directed network, ties *ij*, *jk* and *ik* are observed among actors i, j, and k.

Triangle A Markov graph configuration: in a non-directed network, a clique of three actors, ties *ij*, *jk* and *ik* are observed among actors i, j, and k.

Definition of the Subject

Exponential random graph models, also known as p^* models, constitute a family of statistical models for social networks. These models take the form of a probability distribution of graphs:

$$\Pr(\mathbf{X} = \mathbf{x}) = (1/\kappa)\exp\{\boldsymbol{\eta}'\mathbf{g}\}\,,$$

for a set of tie indicator variables $\mathbf{X}$ on a network of fixed node size n, where $\mathbf{x}$ is a realization, with a parameter vector $\boldsymbol{\eta}$ and a vector of network statistics $\mathbf{g}$. Each value of the parameter vector corresponds to a probability distribution on the set of all graphs with n nodes.

Introduction

As noted in [34], statistical approaches to social networks have quite a long history, stretching at least back to the work of Moreno in the 1930s [24]. Yet, that history is rather sparsely scattered across various literatures and different eras and, as a result, even today we see that older techniques and approaches are reinvented and repackaged as new. The *statistical modeling* of a social network – as distinct from the use of statistically-based approaches to understanding particular properties of a social network – has a shorter and somewhat more coherent pedigree.

Criteria for a successful statistical model of a social network have been proposed by [34]: it should be possible to estimate the parameters of the model from data in a principled way, and the fitted model should be a good representation of that data; the model should provide theoretically plausible interpretations about the type of effects that might have produced the network; and using the estimated model parameters it should be possible to draw inferences about competing explanations for the data. Most, perhaps all, models for social networks fall short of completely meeting these requirements. Some models are "thought experiments", intended to illuminate but with an uncertain link to data. Some models are estimable from data but still cannot adequately represent important features of the network. Models may imply interpretations that are just not theoretically plausible in terms of social science; and some models cannot include different effects simultaneously in order to test one against the other. Of course, by definition all models are imperfect (else, they are not "models"). So these criteria direct the aim in model development, even under the knowledge that we cannot always hit the target exactly.

A major insight in the statistical modeling of social network structure has been that structural effects can be detected as some form of deviation from what would be expected as "randomness". This was in fact the original proposal of Moreno and colleagues [24] who were the first to compare observed network data to what would be expected from null distributions. More explicitly, Rapoport's biased net theory [32,33] proposed certain structural "biases" away from random tie formation, including the important notions of transitivity (later to be described as clustering), the propensity for human social networks to exhibit a high proportion of triangulated ties.

Of course, in the graph theory literature the best-known model for randomness is the *simple random graph model*, often known as the Erdös–Rényi graph [7] or uniform Bernoulli graph distribution. This model proposes that for a given number of nodes network ties are observed between pairs of nodes independently and with a fixed probability p. Properties of this model have been examined extensively, utilizing its analytic tractability. It has also been used as a null model for various sampling distributions (e. g. see the summary in [8].) For an observed network, p can be readily estimated (the maximum likelihood estimate is simply the density of the network), but unfortunately this model is not a good representation of almost any human social network. There are no structural effects here, only randomness.

For directed networks, [15] extended the Bernoulli model in the early 1980s by parametrizing structural effects for reciprocity and for differential node-level activity (out-degree) and popularity (in-degree). Possible network ties within dyads were dependent on one another, but were independent between dyads. This dyad-independence assumption permitted the model to be estimated as a straight-forward loglinear model. Holland and Leinhardt called this model p_1, the subscripted '1' implying a program of further research, with progressively enlarged dependence assumptions (within arcs, within dyads, within triads, etc.). However, Holland and Leinhardt were uncertain how to progress beyond dyads. The problem was that standard statistical estimation required some level of independence and triads, unlike dyads, overlapped. (A subsequent, more sophisticated extension, the p_2 model, also has dyadic independence at its heart, but conditional on random node-level effects, so that the random effects indirectly introduce dependencies that may extend beyond dyads [53].)

Ove Frank and his colleague, David Strauss, provided the crucial insight: in the world of complex networks, it was dependence that mattered, and assumptions involving traditional statistical independence, although helpful for fitting models, were likely to be inadequate empirically [9]. In contrast, their approach centered on *conditional* independence, whereby contingencies among network ties may transmit across the network in the sense that the presence of one network tie may affect the presence of any other, but most pairs of ties were in some sense "remote", so that the contingencies affecting a given network tie had to be transmitted through "neighboring" ties. So once one had a concept of "neighboring ties", that is, some notion of which possible ties were conditionally independent after taking into account other observed network ties, the form of a model could be specified. Approaches from spatial statistics and statistical mechanics, with notions of dependence within neighborhoods (broadly defined) could be translated in this way into social network models. Frank and Strauss proposed Markov random graph models, based on an explicit argument about conditional independence among networks ties, that could be regarded as theoretically plausible for the first time. These models incorporated reciprocity, degree-based effects and various forms of triangulation or clustering.

In the 1990s, Markov random graph models became the accepted form of exponential random graph models and they have only recently been superseded. They could be estimated through the rough and ready procedure of pseudo-likelihood [52], a not particularly sophisticated approach but enough to get the field started empirically. Wasserman and Pattison popularized these models among social networks researchers as p^* models [55] and added further generalizations, with extensions to multivariate [29] and valued networks [36]. Nodal attributes were introduced in social selection [37] and social influence models [38].

In the last decade, especially with more principled methods of estimation and simulation, it has been shown that homogeneous Markov random graph models face considerable, often insurmountable, difficulties in dealing with real network data. Data with inhomogeneities – for instance, very high degree nodes, or dense regions of multiple triangulation, both of which are not uncommon in real social networks – typically result in *degeneracy* for Markov random graph models, where no set of parameter estimates can adequately represent the data. Newer specifications have been proposed by Snijders and colleagues [48] that help considerably with these issues of degeneracy, and have the capacity to enhance our ability to model small-scale to medium-sized social networks, in some cases remarkably well. These advances are based on new theorizations of dependence among network ties, echoing the original project of Holland and Leinhardt with a notion of increasingly extended dependence assumptions.

This article begins with a presentation of notation and terminology and then discusses how dependence hypotheses lead to the general form of the model. Specific examples of dependence hypotheses, and the resulting models, are then presented: Bernoulli graph distributions; dyadic independence models; and Markov random graph models. Methods of simulation are briefly introduced before a discussion of degeneracy issues, particularly as they apply to Markov random graph models. Recently proposed dependence hypotheses are introduced, including the so-called "social circuit" dependence, leading to additional specifications for exponential random graph models that substantially improve model performance. Estimation and goodness of fit approaches are discussed before a short empirical example is presented. Extensions are briefly described followed by some concluding remarks about future directions.

Notation and Terminology

A network comprises a set of relational ties between pairs of individual *actors* (be they people or other social entities). A network can be represented as a graph G with node set $N = \{1, 2, \ldots, n\}$ representing the individuals and edge set E representing the relational ties. For the statistical models of this article, relational ties are construed as a set of binary random variables X_{ij} such that $X_{ij} = 1$ if a tie is observed from node i to node j, for $i \neq j$, and $X_{ij} = 0$, otherwise. For a nondirected network X_{ij} and X_{ji} are equivalent, whereas for a directed network the two variables are distinct. X_{ii} is undefined (or may be considered as a structural zero). A nondirected tie is an *edge* and a directed tie an *arc*. We specify x_{ij} as the observed value of the variable X_{ij} and we let $\mathbf{X}$ be the matrix of all variables with $\mathbf{x}$ being a realization (where the diagonals of the matrices are forced to be zero). $\mathbf{X}$ and $\mathbf{x}$ are necessarily symmetric for nondirected networks, but not so for directed networks. More generally, $\mathbf{x}$ may be valued but in this article we restrict attention to binary ties.

There are natural extensions to this basic notation, depending on the data structures. For instance, suppose the data is in the form of a bipartite network, with two distinct sets of nodes and ties between nodes of different types but not between nodes of the same type (e. g. such a data structure can represent people's membership of clubs). Then $\mathbf{X}$ may comprise a set of variables X_{pa} indicating the presence or absence of a tie between a node p of the first type and a node a of the second type. The data may involve *multiple* or *multivariate networks*, with r different types of relations among the one set of nodes. In that case $\mathbf{X}$ may be thought of as a three-way array of variables X_{ijr} such that $X_{ijr} = 1$ indicates that presence of a tie of type r from node i to node j. An analogous three-way array may represent network data collected for the same set of nodes at different time points such that $X_{ijt} = 1$ indicates that presence of a tie at time t from node i to node j.

Nodes themselves can have certain properties (*attributes*) that may be measured as binary, categorical or continuous variables. We denote an attribute variable with the vector $\mathbf{Y}$, where $Y_i = y_i$ indicates that for attribute Y node i has a value y_i.

Dependence Hypotheses

Following work in spatial statistics [1], Frank and Strauss introduced the notion of a *dependence graph* into social network modeling [9] in order to represent possible dependencies among network variables X_{ij}. The nodes of the dependence graph are the network variables X_{ij} and an edge between two nodes indicates dependence between the respective variables, even when the observed values for all remaining variables are known. These edges specify a *neighborhood* relationship between pairs of variables and the cliques of the dependence graph can be thought of as *local social neighborhoods* for the set of tie variables [27]. Conversely, the absence of an edge in the dependence graph indicates that two network variables are conditionally independent, and are not neighbors, so that their interaction need not be taken into account in a model based on local social neighborhoods. (A more technical discussion on dependence graphs for social network models can be found in [35]).

The previous paragraph describes one important result from the Hammersley–Clifford theorem [1], that the form of a probability distribution for a set of interacting variables relates solely to the neighborhood structure (or clique structure) of the dependence graph (with a single network variable also taken as a clique). For a given node set N, once a dependence hypothesis is specified and the dependence graph is defined, it follows *necessarily* from the Hammersley–Clifford theorem that:

$$\Pr(\mathbf{X} = \mathbf{x}) = (1/\kappa(\boldsymbol{\eta})) \exp\{\Sigma_A \eta_A g_A(\mathbf{x})\} , \qquad (1)$$

where:

(i) the summation is over all neighborhoods A;
(ii) η_A is a parameter corresponding to the neighborhood A (and there is no non-zero parameter for any set of ties that do not constitute a neighborhood);
(iii) $g_A(\mathbf{x}) = \prod_{x_{ij} \in A} x_{ij}$ is the *network statistic* corresponding to neighborhood A and indicates whether all the ties in A are observed in the network $\mathbf{x}$;
(iv) κ is a normalizing quantity, that is a function of the parameter vector $\boldsymbol{\eta}$, to ensure that (1) is a proper probability distribution.

All exponential random graph models take this general form, describing a probability distribution of graphs on n nodes. Neighborhoods A are subsets of possible ties, so they represent possible subgraphs that may or may not be observed in the network $\mathbf{x}$. There is one, and no more than one, parameter for each distinct neighborhood (although one neighborhood may be a subgraph of another and so both may have separate parameters).

The form of the model in (1) is too general to be identifiable and some homogeneity constraint needs to be imposed on what is otherwise a large number of neighborhoods A (and hence a large number of parameters). The original proposal [9] was that parameters should be equated across subsets of ties that were isomorphic to each other (that is, subgraphs that are indistinguishable once node labels are removed). With some further constraints (noted below), this assumption can produce identifiable models that can be fitted to data. (Other possible ways to constrain the number of parameters are also described below). Following the terminology first used by Moreno [24], we term these isomorphic neighborhoods as *network configurations* (although some of the network literature of the last decade has adopted the term *motif*). Counts of configurations become the sufficient statistics of the model.

Examination of the parameters for the various configurations of a fitted model can provide insight into the social processes that may underpin the network. The strength and direction of any particular parameter value will affect how frequently the corresponding configuration is observed. If the parameter is large and positive, we expect to observe the corresponding configuration to occur more frequently than if the parameter value were zero [40].

Obviously, good dependence hypotheses are necessary for this approach to work, as they crucially shape the nature of the model by defining the possible configurations.

Bernoulli Random Graph (Erdös–Rényi) Models

Suppose we hypothesize that there are no dependencies within the network, so that the dependence graph has no edges and the only neighborhoods are single tie variables X_{ij}. Then (1) becomes:

$$\Pr(\mathbf{X} = \mathbf{x}) = (1/\kappa) \exp(\Sigma_{i,j} \eta_{ij} x_{ij}) .$$

Constraining parameters to be equal for isomorphic neighborhoods here implies equating all the parameters $\eta_{ij} = \theta$, as there is only one type of neighborhood. Then we have:

$$\Pr(\mathbf{X} = \mathbf{x}) = (1/\kappa) \exp(\Sigma_{i,j} \theta x_{ij}) = (1/\kappa) \exp(\theta L(\mathbf{x})) ,$$

where $L(\mathbf{x})$ is the number of edges (arcs) in the network $\mathbf{x}$ and θ is a *density* or *edge* parameter (sometimes called a *choice* parameter in older literature). This graph distribution is equivalent to that of a distribution of simple random graphs where the probability of a tie between a pair of edges is $\exp\theta/(1 + \exp\theta)$.

The dependence assumption here is unrealistic and this model will fit few, if any, real networks, but it is often a useful null or baseline model.

Dyadic Independence Models

For directed networks, a dyadic independence hypothesis implies edges in the dependence graph between variables

X_{ij} and X_{ji}. Neighborhoods then take the form of single edges and dyadic pairs $\{X_{ij}, X_{ji}\}$. Applying constraints in the form of the one density parameter θ to all single edge neighborhoods and the one *reciprocity parameter* ρ to all dyadic neighborhoods results in the following two parameter model:

$$\Pr(\mathbf{X} = \mathbf{x}) = (1/\kappa) \exp(\theta L(\mathbf{x}) + \rho M(\mathbf{x}))\,,$$

where $L(\mathbf{x})$ is the number of arcs and $M(\mathbf{x})$ the number of mutual dyads (i. e. where $X_{ij} = X_{ji} = 1$) in the network $\mathbf{x}$. This model has the capacity to model reciprocity effects in directed networks. Slightly less sweeping homogeneity constraints result in the p_1 model [15].

The interpretation of a single parameter in an exponential random graph model is relative to other effects in the model, so that in this case, the presence of a large and positive ρ parameter is relative to the density effect. This is an advantage: in this model, for instance, we can infer that there is substantial reciprocity *given the density of the graph*. It is of little use to talk about a large reciprocity effect in absolute terms, because the number of arcs in the graph (i. e. the density) establishes the preconditions for reciprocity. In other words, graphs of very low density have very little opportunity to demonstrate reciprocity anyway, and the absence of mutual dyads in a low density graph may simply be a feature of the low density. Similarly the presence of many mutual dyads in a high density graph may simply be explained by the high density, without the need to infer a reciprocity effect. A large positive ρ, on the other hand, shows that the number of arcs in a graph are arranged in sufficiently many mutual dyads to suggest that a separate and substantial reciprocity process is required to explain the structure of this network, over and above any density effect.

Once again, the dyadic independence assumption does not do a good job of reproducing real networks, because the model has no capacity for triangulation. The model may be useful as a baseline model.

Markov Random Graphs

Frank and Strauss [9] proposed Markov dependence, by postulating that a possible tie from i to j is assumed to be contingent on any other possible tie involving i or j, even if all other ties in the network are fixed. Markov dependence implies that two possible network ties are conditionally independent unless they share a common actor. They showed that this assumption resulted in models with configurations for nondirected graphs of single edges, star-like structures and triangles (or 3-cycles). For directed graph models, configurations include directed counterparts of these as well as reciprocity (see [55] for a fuller description of directed graph parameters). Frank and Strauss termed these *Markov random graph models*.

Once homogeneity is imposed across isomorphic neighborhoods, we obtain configurations (and related parameters) for nondirected networks as depicted in Fig. 1. The resulting model is:

$$\Pr(\mathbf{X} = \mathbf{x}) = (1/\kappa) \exp(\theta L(\mathbf{x}) + \Sigma_{r=2,n-1}\sigma_r S_r(\mathbf{x}) + \tau T(\mathbf{x}))\,, \quad (2)$$

where:

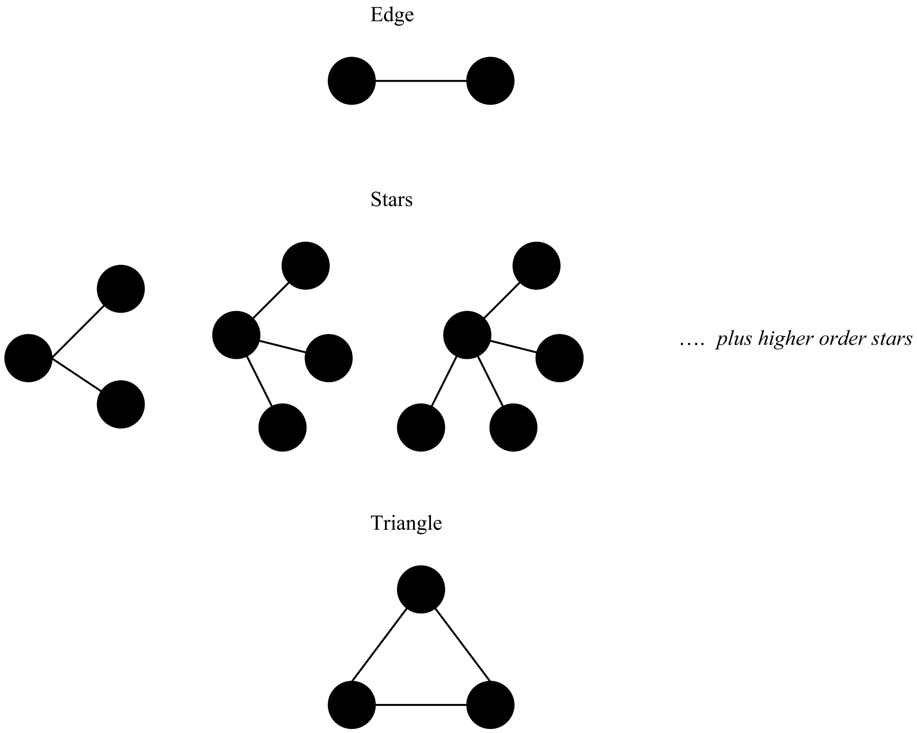
(i) θ is a density parameter, and $L(\mathbf{x})$ the number of edges in $\mathbf{x}$, as before;
(ii) σ_r is a parameter for a *star of size* r, and $S_r(\mathbf{x})$ is the number of stars of size r in $\mathbf{x}$;
(iii) a star of size r is a configuration centered on a single node i such that there are r edges emanating from i (note that stars of size larger than r contain many stars of size r, so that this is not simply a partition into nodes of different degrees, although the full set of counts $S_r(\mathbf{x})$ can be converted into the degree distribution, and vice versa);
(iv) τ is a triangle or clustering parameter and $T(\mathbf{x})$ is a count of the number of triangles in $\mathbf{x}$.

This model still has too many star parameters to be identifiable. Note that the expression $\theta L(\mathbf{x}) + \Sigma_{r=2,n-1}\sigma_r S_r(\mathbf{x})$ completely parametrizes the degree distribution (so the τ parameter can be interpreted as the strength of transitivity conditional on the degree distribution of the graph). An identifiable model can be obtained by restricting the number of star parameters to be substantially less than $n - 2$, for instance, to less than four (an alternative approach is described below). Then the model becomes:

$$\Pr(\mathbf{X} = \mathbf{x}) = (1/\kappa) \exp(\theta L(\mathbf{x}) + \sigma_2 S_2(\mathbf{x}) + \sigma_3 S_3(\mathbf{x}) + \tau T(\mathbf{x}))\,. \quad (3)$$

The parameters can be interpreted in ways that represent plausible social processes [39]:

(i) θ is a baseline propensity for social ties to form;
(ii) σ_2 represents a tendency for individuals to seek multiple partners (if it is positive);
(iii) whereas σ_3 (if negative as it often is when this model is fitted to real data) represents a ceiling effect against having too many partners; so, together the balance between σ_2 and σ_3 represents the benefits of multiple



Social Networks, Exponential Random Graph (p^*) Models for, Figure 1
Configurations for a nondirected Markov random graph model

contacts against the costs of maintaining too many contacts;

(iv) while τ represents tendencies towards clustering.

An alternative interpretation is to consider the θ, σ_2 and σ_3 parameters as controlling for the first three moments of the degree distribution, so that while model (2) in effect presents a full parametrization, model (3) provides a more parsimonious control over the degree distribution (saturated models of the degree distribution are discussed by [43] and [47], see also [11]). Of course, for any data set it is an empirical question whether the first three moments are sufficient to capture the degree distribution adequately. Below, we describe ways to investigate that question.

Extensions of this basic Markov random graph model include models for multivariate networks [29], for valued networks [36] and for affiliation networks [42] (see also [28]).

By tuning parameter values even the simple four parameter model (3) can represent many different types of networks, including small world networks, networks with long paths, and highly clustered ("caveman") networks [39]. For some parameter values, however, the models became "frozen" into certain unchanging highly structured patterns. This effect is an illustration of *degeneracy* for these models, which was a research issue given important attention in the early years of the 21st century. The upshot of this body of research was that homogeneous Markov random graph models were recognized as frequently inadequate to deal with real social network data.

Simulation and Model Degeneracy

The problems of degeneracy became apparent through a number of simulation studies. Strauss was the first to describe a rather straightforward application of standard statistical simulation techniques (e. g. Metropolis or Metropolis–Hastings algorithms) to simulate Markov random graph models for a fixed set of nodes and given set of parameter values [51]. More recent treatments and some important results are given in [13,14,17,39,45]. These methods provide a principled statistical means to produce a distribution of graphs with probabilities of each graph being observed in the distribution consistent with any particular model in the form of (1). Typically the simulation is run for a large number of iterations and a sample of graphs is extracted and examined to understand typical properties of the graphs in the distribution.

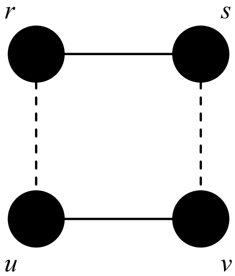
A graph distribution is termed as *near degenerate* [13,14] if it implies only a very few (possibly only one or two) distinct graphs with substantial non-zero probabilities. For instance, certain parameter values for Markov random graph models place almost all of the probability mass on either the empty or the full graph. These cannot be good models for human social networks. For certain parameter values, there may be two quite separate regions where graphs are most likely (possibly one region of low density and another of high density graphs), with the possibility of a dramatic phase transition from one region to the other (for instance, as parameter values change very slightly). Such phase transitions and other aspects of degeneracy have been studied for a variety of simple Markov random graph models by a number of authors [2,12,13,14,19,26,39,41,45,48].

Issues of degeneracy call into question whether Markov random graph models can adequately represent most network data. It would not matter if some regions of the parameter space produced degenerate models, as long as those regions applicable to empirical data were non-degenerate. Experience shows the opposite applies: for some data, Markov random graph models turn out to be well-behaved, but for many empirical networks they are degenerate. When applied to social networks with a few very high degree nodes, or with some regions of high triangulation, degeneracy frequently occurs for Markov models. Models such as (3) reflect homogeneity constraints, and they appear to have difficulty when the data presents inhomogeneity in the distribution of degrees or triangles. It is not entirely certain whether the problem with the models arises because of the Markov dependence assumption or the assumption of homogeneity. In any event, to deal more effectively with real networks, a different approach to these models seems necessary. To date, attention has been directed to revisiting the dependence assumption.

Snidjers and colleagues [48] drew two conclusions from the tendency for Markov random graph models to be degenerate when fitted to real networks with a high level of clustering: (1) the Markov dependence assumption may be too restrictive; (2) the representation of the social phenomenon of transitivity by the total number of triangles might be too simplistic. They proposed specifications that drew on both these possibilities. Before we introduce these new specifications, however, we turn to higher order network dependencies.

Social Circuit Dependence: Partial Conditional Dependence Hypotheses

There are several ways in which network dependence assumptions could be extended beyond Markov dependence [27]. One approach is of *partial conditional dependence*, whereby the presence of certain network ties *created* dependence among other possible network ties. This assumption permits the emergence of dependence structures as ties come into and go out of existence. This type of dependence can be incorporated into a version of the Hammersley–Clifford theorem and hence into exponential random graph models [27].



Social Networks, Exponential Random Graph (p^*) Models for, Figure 2
Social circuit dependence (*Broken lines* represent possible edges; *full lines* represent observed edges X_{ru} is conditionally dependent on X_{sv})

A particular hypothesis about this type of emergent dependence was used by [48] and described by [41] as *social circuit dependence*. Social circuit dependence is defined as two possible network ties being conditionally dependent if their observation would lead to a 4-cycle. This type of dependence is depicted in Fig. 2. Here $X_{rs} = X_{uv} = 1$, that is, in the network data there are observed edges between nodes r and s, and between nodes u and v. The presence of these edges leads to conditional dependence between variables X_{ru} and X_{sv} because if edges were also observed between r and u, and between s and v, a 4-cycle would result in the graph.

Interpretation of this hypothesis is discussed by [41]. Generally we would expect two distinct possible edges (r, u) and (s, v) to be conditionally independent. If, however, person r knows person s, and person u knows person v, then the presence of a tie between r and u can make the presence of a tie between s and v more likely, that is, they are conditionally dependent. It is simple to think of real social circumstances where this happens: for instance, in families the presence of a friendship between two children increases the chances of the two mothers coming to know one another; or in a business, cooperation between two bosses may lead to their employees working together; or in research, two postdoctoral researchers might discuss issues with each other because their academic mentors have been collaborators. In all of these cases, what makes the dependence come into effect is the presence of ties that constitute part of the 4-cycle. Without those ties, the dependence does not arise. So, the mothers are the mothers of the two particular children, not the mothers of entirely different children; the employees are employed by those particular bosses; and the two post-docs are supported by those two mentors.

Social Circuit Specifications

Snijders and colleagues [48] proposed three new statistics for exponential random graph models for nondirected networks: *alternating k-stars*, *alternating k-triangles*, and *alternating independent two-paths*. To derive these statistics they use: (1) social circuit dependence *in addition to* Markov dependence; and (2) a non-linear functional form combining various configurations into the one statistic.

Alternating *k*-Stars

The alternating k-star parameter uses the second idea alone and does not extend dependence beyond the class of Markov random graph models. Rather than adopt the usual practice of limiting the number of higher order stars – as for example in model (3) where the star parameters are deliberately limited to no more than 3-stars – Snijders and colleagues returned to model (2), keeping all σ_r in the model but imposing constraints among these parameter values. Specifically, the *alternating k-star assumption* proposes that for all $k \geq 2, \sigma_{(k+1)} = -\sigma_k/\lambda$ for some λ greater than 1. Then in (1) there is one parameter σ for all star effects, with an associated statistic:

$$u = \sum_{k=2}^{n-1} (-1)^k \frac{S_k}{\lambda^{k-2}} , \tag{4}$$

where the parameter σ is referred to as the *alternating k-star parameter*. As λ is greater than 1, the direct impact of higher order stars (that is, very high degree nodes) is reduced for higher k. Of course, as noted above, high degree nodes still have substantial effect in the model as they produce many 2- and 3-stars, and so on. The alternating sign helps balance these overlapping star counts for high degree nodes. In [48] λ is set at 2 but [17] show how to estimate an optimal value of λ (see also [16]). The alternating k-star parameter is equivalent to a *geometrically weighted degree parameter* that explicitly models the degree distribution but puts more weight on the numbers of nodes with lower degrees, with weights decreasing geometrically as the degrees increase [16].

The interpretation of the parameter is as follows [41, 48]. If the alternating k-star parameter is positive, then highly probable networks are likely to contain some higher degree nodes. A positive alternating k-star parameter (together with a negative density parameter) implies graphs that exhibit preference for connections between a larger number of low degree nodes and a smaller number of higher degree nodes, akin to a core-periphery structure.

Alternating *k*-Triangles

The alternating k-triangle assumption incorporates both social circuit and Markov dependence. Snijders and colleagues show that with both these dependence assumptions operating simultaneously, configurations more complex than simple triangles are possible [48]. In particular, they introduced the notion of a *k-triangle*, a combination of k individual triangles that all share one edge (the common *base* of the k triangles), as represented in Fig. 3.

Let T_k be the count of k-triangles in a graph. Then the *alternating k-triangles assumption* combines these counts into the one statistic in an analogous way as for the alternating k-stars, that is:

$$t = 3T_1 + \sum_{k=1}^{n-3} (-1)^k \frac{T_{k+1}}{\lambda^k} , \tag{5}$$

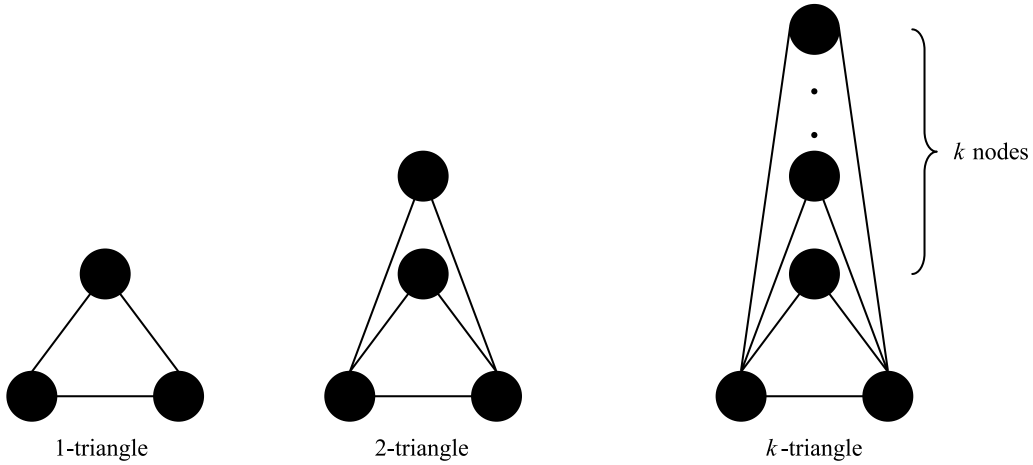
where the factor of 3 for T_1 is due to symmetry considerations in a non-directed triangle [48].

This is the *alternating k-triangle statistic* with an associated *alternating k-triangle parameter* τ. A positive k-triangle parameter indicates triangulation in the network but also tendencies for triangles themselves group together in larger higher order "clumps". A positive alternating k-triangle effect combined with a negative alternating k-star effect can produce a web of multiple smaller regions of triangulation [41,48].

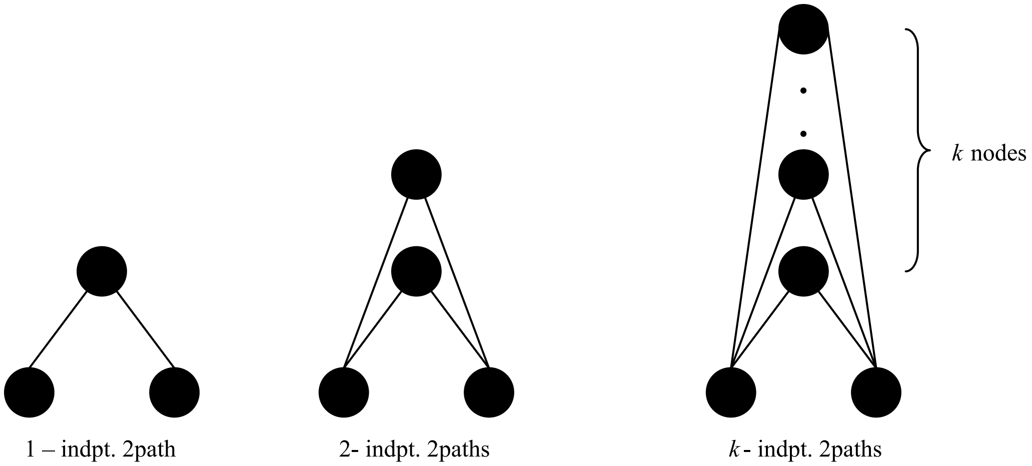
[16] shows that the alternating k-triangle parameter is equivalent to the *edge-wise shared partner* (or ESP) parameter that models the distribution of shared partners of tied actors, but with weights decreasing geometrically as the number of shared partners increase.

Alternating *k*-Two-Paths

Snijders and colleagues also proposed a parameter that represents a lower order configuration for a k-triangle, namely a *k-two-path* which is a k-triangle without the base [48]. These configurations, represented in Fig. 4, describe the number of distinct two-paths between a pair of nodes. The motivation was to provide a parameter that, in conjunction with k-triangles, would distinguish between tendencies to form edges at the base, or at the sides of a k-triangle.



Social Networks, Exponential Random Graph (p^*) Models for, Figure 3
Various *k*-triangles



Social Networks, Exponential Random Graph (p^*) Models for, Figure 4
Various alternating *k*-independent 2path configurations

Again the counts of these configurations are combined into the one alternating statistic in an analogous way as in the previous two cases, to produce an alternating 2-path statistic and parameter for the model. [16] shows that this new parameter is equivalent to the *dyad-wise shared partner* (or DSP) parameter that models the distribution of shared partners of actors who may or may not be tied, but with weights decreasing geometrically as the number of shared partners increase.

The parameter can be interpreted as representing localized multiple connectivity between nodes. When this parameter is negative, together with a positive alternating *k*-triangle parameter, there is a tendency against 4-cycles in the network, unless those cycles include triangles (alternatively, the presence of many 2-paths between nodes is related to the formation of triangles).

Specifications for Directed Graphs

Counterpart statistics and parameters for directed graphs have also been proposed. We do not discuss these here but refer readers to [48].

Estimation

Pseudo-likelihood estimation [52], based on logistic regression techniques, was commonly used until more principled methods became available in recent years. Preferred methods of estimation involve simulation procedures (Monte Carlo Markov chain maximum likelihood estimation – MCMCMLE) which for these models has been discussed by a number of authors [4,5,6,17,45,47,56]. Software for Monte Carlo estimation techniques has recently become publicly available ([41] provides a review). The observed graph statistics are compared with those expected under a set of provisional parameter values, using a stochastic simulation based on those values. The goal is to find a set of parameter values whereby the observed statistics are equal to the expected statistics. Parameter values are refined until the observed and expected statistics can be equated, when the model is said to converge. If the parameter estimates never converge, the model is degenerate. Standard errors can also be estimated.

Goodness of Fit and Comparisons with Markov Models

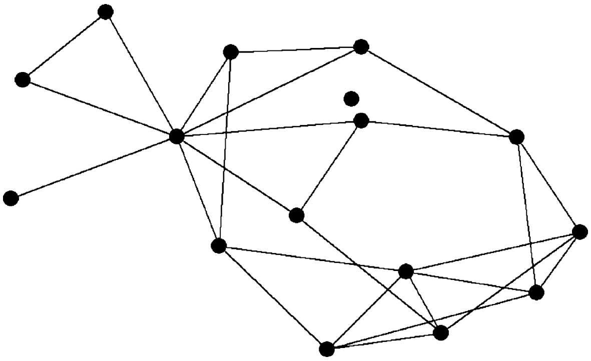
Simulation also provides an innovative means to assess how well the model can represent the data. By simulating the model from the parameter estimates, and extracting a sample of graphs it is possible to examine any graph statistic of interest (whether there is an associated parameter in the model or not). Any graph statistic for the observed graph can be compared to the distribution of such graph statistics from the simulated graphs. If the observed graph statistic is not extreme in the distribution of simulated statistics, then it is plausible that that particular feature of the observed graph could have come from the distribution of graphs implied by the model. In this way, it is possible to check which graph features the model successfully captures. This procedure is described in detail by [18] (see also [41]).

[11] provides a compelling example of how this procedure may be used to make decisions about model selection. On the basis of applying the new parametrization to a large school-based network of over 1000 nodes, [11] concludes that for this data Markov models were degenerate, but the newer parametrization had similar underlying interpretations, avoided degeneracy and was empirically validated. Other comparisons with Markov models demonstrate substantial improvements for the new parametrization in overcoming degeneracy [41].

A Simple Empirical Example

Figure 5 presents an empirical network of social relationships among 16 actors, in this case families from medieval Florence [25].

Markov models do not converge for this network. Table 1 provides parameter estimates and standard errors for models involving the social circuit parameters. The third column in the table provides a convergence statistic which indicates good convergence if it has an absolute value less than 0.1 [41]. In this case, the model exhibits excellent convergence.



Social Networks, Exponential Random Graph (p^*) Models for, Figure 5
Social relationships among 16 actors

Social Networks, Exponential Random Graph (p^*) Models for, Table 1
Parameter estimates: Florentine families network (NB: $\lambda = 2$)

Parameter	Estimate	Standard error	Convergence
Edge	−0.04	2.09	0.01
Alternating *k*-star	−1.01	0.78	0.02
Alternating *k*-triangle	0.68	0.33	0.02
Alternating *k*-2paths	0.18	0.17	0.03

A parameter may be inferred as important if its estimate in absolute terms is more than twice its standard error. So in this simple model, there is a substantial effect for alternating *k*-triangles. In Fig. 5, regions of triangulation can be seen, consistent with the alternating *k*-triangle effect, perhaps combined with the negative alternating *k*-star effect (although the reliability of that negative effect is not certain, given that the estimate is not more than twice the standard error). In sum, there is good evidence here that this network differs substantially from a simple random network, in particular, that there is substantially more triangulation in this network than would be expected in a simple random graph.

Table 2 shows the goodness of fit analysis for graph features other than those in the model. This analysis was based on extracting 1,000 sampled graphs from a simulation of 1,000,000 (after a burn-in of 100,000), using the parameter estimates in Table 1. The goodness of fit statistic takes the form of a *t*-ratio, calculated as the difference between the graph statistic and the mean from the

Social Networks, Exponential Random Graph (p^*) Models for, Table 2
Goodness of fit analysis: Florentine families network model

Graph feature	Goodness of fit
Number of 2-stars	−0.10
Number of 3-stars	−0.25
Number of triangles	−0.17
Standard deviation, degree distribution	−0.10
Skew, degree distribution	−0.12
Number of isolates	1.06
Clustering coefficient	−0.09

simulated graphs divided by the standard deviation as estimated from the sample. For graph features not in the model, the model is considered to reproduce the data well if the goodness of fit statistic is not extreme (for instance, with t-ratios of less than 2). As can be seen, on the features examined, including counts of various Markov configurations, aspects of the degree distribution, and clustering, the model reproduces the data well. There is one isolate in the data, whereas 71% of simulated graphs had no isolates, hence the rather larger goodness of fit statistic. Nevertheless, with 29% of graphs in the sample having at least one isolate, the data cannot be said to be extreme in this regard.

The clustering coefficient in the Table refers to the proportion of actual to possible triangles and is calculated as thrice the number of triangles as a ratio of the number of two-paths (the factor of three applies because there are three two-paths in any one triangle).

Table 2 does not show features of the geodesic distribution but a separate analysis reveals that the observed geodesic distribution is consistent with those from the simulated graphs: none of the geodesic quartiles of the observed graph is extreme compared to the distribution of quartiles from the simulated graphs. Overall, the model can be said to fit the data well.

Further Extensions and Future Directions

Multiple Networks

It is possible to use the same general framework as (1) when $\mathbf{X}$ is a three-way array representing multiple relations on the one set of nodes. The parametrization, however, becomes more complex. [29] showed how to develop Markov models along these lines, with these models subsequently used in some interesting empirical work, principally in organizations [22,23]. This empirical research, however, utilized pseudo-likelihood estimation and did not take into account possible model degeneracy. See also [20].

Early work on adapting the new specifications proposed by Snijders and colleagues to multiple networks has concentrated on the case of two networks, by using the alternating k-star, -triangle, and -2path parameters for effects within networks and then investigating dyadic associations between networks [57].

Bipartite Networks

Again (1) may be used as a basis for models for bipartite networks. [42] were the first to do this with Markov dependence. [28] proposed additional parameters that derived from partial conditional dependence assumptions. [54] incorporated these and other parameters, combined with an alternating combination of configuration counts analogous to those in the newer specifications, into a model for bipartite networks.

Longitudinal Networks

Perhaps the major approach to modeling network data collected at multiple time points is the actor-oriented models of Snijders and colleagues [44] which utilize a continuous time Markov chain technique. Certain specifications of actor-oriented dynamic models have exponential random graph models as their stationary distribution [44], so the links between the two approaches are quite strong. Accordingly, it is possible to develop longitudinal versions of exponential random graph models to investigate network dynamics. The conceptual difference with the actor-oriented approach is quite subtle: exponential random graph longitudinal models suppose that ties change in response to particular social neighborhoods of other ties, rather than in response to actors seeking to optimize particular structural positions. Early work along these lines, including the evolution of multiple networks, has been reported by [30]. For further discussion on the links between the actor-oriented and tie-based versions of the models, see [46].

Nodal Attributes

One of the important issues that researchers often wish to examine is whether and how nodal attributes relate to network structure. For instance, there are good empirical and theoretical grounds for expecting that people who have social relationships are more likely to be similar to each other (known as *homophily*). Conceptually, two possible processes apply: individuals may develop ties because they are similar (*social selection*) or individuals who are tied may influence each other to be similar (*social influence*).

Of course, both of these processes may occur simultaneously and it is a methodological challenge for a given set of data to disentangle the two to determine the respective strengths of selection and influence effects, respectively. Snijders and colleagues [31,49,50] have recently developed innovative methods for investigating this question for longitudinal data within actor-oriented longitudinal models.

[37] proposed social selection models within the framework of (1), having the form:

$$\Pr(\mathbf{X} = \mathbf{x}|\mathbf{Y} = \mathbf{y}) = (1/\kappa)\exp\{\Sigma_A \eta_A g_A(\mathbf{x}, \mathbf{y})\}\,, \qquad (6)$$

where **Y** is a vector of attribute variables (binary, categorical or continuous). Here the configurations A relate to combinations of nodal attributes and network subgraphs, and the network is modeled conditional on a fixed distribution of attributes. The intent of such models may be various: accounting for heterogeneity (that may otherwise cause difficulties in fitting models) as well as assessing selection, homophily, and covariate activity and popularity, while controlling for dependencies naturally occurring in the network setup.

An analogous approach to social influence models and network contagion was proposed by [38] where the patterns of attributes were modeled conditional on a fixed network:

$$\Pr(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x}) = (1/\kappa)\exp\{\Sigma_A \eta_A g_A(\mathbf{x}, \mathbf{y})\}\,. \qquad (7)$$

Future Directions

Not only have the new specifications shown a remarkable improvement in the successful modeling of real networks, they have opened up a wide range of network modeling possibilities. The discussion of model extensions mentioned above gives some indication of current work in this area. Of course, the possibility of further improvements in model specification has not been closed. Future directions may include methods to examine very large, community-based networks, not only in terms of data collection (for instance, the development of model-based sampling methods) but also theoretical issues based on possible differences in dependence between very large scale and smaller scale structures. Model-based approaches to missing network data, perhaps within a Bayesian framework, are also in prospect [10,21].

The use of exponential random graph models in empirical research, especially in combination with multiple measures on individuals, opens a wide prospect for network statistical modeling to make a substantial contribution in many applied contexts.

Bibliography

Primary Literature

1. Besag J (1974) Spatial interaction and the statistical analysis of lattice systems. J R Stat Soc Ser B 36:96–127
2. Burda Z, Jurkiewicz J, Kryzwicki A (2004) Network transitivity and matrix models. Phys Rev E69:026106
3. Corander J, Dahmström K, Dahmström P (1998) Maximum likelihood estimation for Markov graphs, Research report. Department of Statistics, Stockholm University, Stockholm
4. Corander J, Dahmström K, Dahmström P (2002) Maximum likelihood estimation for exponential random graph models. In: Hagberg J (ed) Contributions to social network analysis, information theory and other topics in statistics: A festschrift in honour of Ove Frank. Department of Statistics, Stockholm University, Stockholm
5. Crouch B, Wasserman S (1998) Fitting p^*: Monte Carlo maximum likelihood estimation. International Conference on Social Networks, Sitges, Spain, May 28–31
6. Dahmström K, Dahmström P (1993) ML-estimation of the clustering parameter in a Markov graph model, Research report. Department of Statistics, Stockholm University, Stockholm
7. Erdös P, Rényi A (1959) On random graphs. I. Publicationes Mathematicae (Debrecen) 6:290–297
8. Frank O (1981) A Survey of Statistical Methods for Graph Analysis. Sociol Methodol 12:110–155
9. Frank O, Strauss D (1986) Markov graphs. J Am Stat Assoc 81:832–842
10. Gile K, Handcock M (2006) Model-based assessment of the impact of missing data on inference for networks. University of Washington, CSS working paper 66
11. Goodreau S (2007) Advanced in exponential random graph (p^*) models applied to a large social network. Soc Netw 29:231–248
12. Häggström O, Jonasson J (1999) Phase transition in the random triangle model. J Appl Probab 36:1101–1115
13. Handcock M (2002) Statistical models for social networks: Degeneracy and inference. In: Breiger R, Carley K, Pattison P (eds) Dynamic social network modeling and analysis. National Academies Press, Washington, pp 229–240
14. Handcock M (2003) Assessing degeneracy in statistical models of social networks. University of Washington CSS Working Paper no 39
15. Holland P, Leinhardt S (1981) An exponential family of probability distributions for directed graphs (with discussion). J Am Stat Assoc 76:33–65
16. Hunter D (2007) Curved exponential family models for social networks. Soc Netw 29:216–230
17. Hunter D, Handcock M (2006) Inference in curved exponential families for networks. J Comput Graph Stat 15:565–583
18. Hunter D, Goodreau S, Handcock M (2008) Goodness of fit of social network models. J Am Stat Assoc 103:248–258
19. Jonasson J (1999) The random triangle model. J Appl Probab 36:852–867
20. Koehly L, Pattison P (2005) Random graph models for social networks: multiple relations or multiple raters. In: Carrington P, Scott J, Wasserman S (eds) Models and methods in social network analysis. Cambridge University Press, New York, pp 162–191

21. Koskinen J (2004) Essays on Bayesian inference for social networks. Department of statistics, Stockholm University, Stockholm
22. Lazega E, Pattison P (1999) Multiplexity, generalized exchange and cooperation in organizations. Soc Netw 21:67–90
23. Lomi A, Pattison P (2006) Manufacturing relations: An empirical study of the organization of production across multiple networks. Organ Sci 17:313–332
24. Moreno J, Jennings H (1938) Statistics of social configurations. Sociometry 1:342–374
25. Padgett JF, Ansell CK (1993) Robust action and the rise of the Medici, 1400-1434. Am J Sociol 98:1259–1319
26. Park J, Newman M (2004) Solution of the 2-star model of a network. Phys Rev E70:066146
27. Pattison P, Robins G (2002) Neighborhood-based models for social networks. Sociol Methodol 32:301–337
28. Pattison P, Robins G (2004) Building models for social space: Neighbourhood based models for social networks and affiliation structures. Mathematiques des science humaines 168:11–29
29. Pattison P, Wasserman S (1999) Logit models and logistic regressions for social networks, II. Multivariate relations. Br J Math Stat Psychol 52:169–194
30. Pattison P, Robins G, Wang P, Snijders TAB, Koskinen J (2006) The co-evolution of multiple networks. Sunbelt XXVI International Social Networks Conference, Vancouver, April 2006
31. Pearson M, Steglich C, Snijders TAB (2006) Homophily and assimilation among sport-active adolescent substance users. Connections 27:51–67
32. Rapoport A (1953) Spread of information through a population with a socio-structural bias: I. assumption of transitivity. Bull Math Biophys 15:523–533
33. Rapoport A (1957) Contributions to the theory of random and biased nets. Bull Math Biophys 19:257–277
34. Robins G, Morris M (2007) Advances in exponential random graph (p^*) models. Soc Netw 29:169–172
35. Robins G, Pattison P (2005) Interdependencies and social processes: Generalized dependence structures. In: Carrington P, Scott J, Wasserman S (eds) Models and Methods in Social Network Analysis. Cambridge University Press, New York, pp 192–214
36. Robins G, Pattison P, Wasserman S (1999) Logit models and logistic regressions for social networks, III. Valued relations. Psychometrika 64:371–394
37. Robins G, Elliott P, Pattison P (2001) Network models for social selection processes. Soc Netw 23:1–30
38. Robins G, Pattison P, Elliott (2001) Network models for social influence processes. Psychometrika 66:161–190
39. Robins G, Pattison P, Woolcock J (2005) Social networks and small worlds. Am J Sociol 110:894–936
40. Robins G, Pattison, Kalish Y, Lusher D (2007) An introduction to exponential random graph (p^*) models for social networks. Soc Netw 29:173–191
41. Robins G, Snijders TAB, Wang P, Handcock M, Pattison P (2007) Recent developments in exponential random graph (p^*) models for social networks. Soc Netw 29:192–215
42. Skvoretz J, Faust K (1999) Logit models for affiliation networks. Sociol Methodol 29:253–280
43. Snijders TAB (1991) Enumeration and simulation methods for 0–1 matrices with given marginals. Psychometrika 56:397–417
44. Snijders TAB (2001) The statistical evaluation of social network dynamics. Sociol Methodol 31:361–395
45. Snijders TAB (2002) Markov chain Monte Carlo estimation of exponential random graph models. J Soc Struct 3:2
46. Snijders TAB (2006) Statistical methods for network dynamics. In: Luchini S et al (eds) Proceedings of the XLIII Scientific Meeting, Italian Statistical Society, pp 281–296
47. Snijders TAB, van Duijn MAJ (2002) Conditional maximum likelihood estimation under various specifications of exponential random graph models. In: Hagberg J (ed) Contributions to social network analysis, information theory, and other topics: A festschrift in honour of Ove Frank. Department of Statistics, University of Stockholm, Stockholm, pp 117–134
48. Snijders TAB, Pattison P, Robins G, Handcock M (2006) New specifications for exponential random graph models. Sociol Methodol 36:99–153
49. Snijders TAB, Steglich C, Schweinberger M (2007) Modeling the co-evolution of networks and behavior. In: van Montfort K, Oud H, Satorra A (eds) Longitudinal models in the behavioral and related sciences. Lawrence Erlbaum, Hillsdale, pp 41–71
50. Steglich C, Snijders TAB, West P (2006) Applying SIENA: An illustrative analysis of the coevolution of adolescents' friendship networks, taste in music, and alcolhol consumption. Methodology 2:48–56
51. Strauss D (1986) On a general class of models for interaction. SIAM Rev 28:513–527
52. Strauss D, Ikeda M (1990) Pseudo-likelihood estimation for social networks. J Am Stat Assoc 85:204–212
53. van Duijn M, Snijders TAB, Zijlstra B (2004) p2: a random effects model with covariates for directed graphs. Stat Neerlandica 58:234–254
54. Wang P (2006) Exponential random graph (p^*) models for affiliation networks. Postgraduate diploma thesis, University of Melbourne
55. Wasserman S, Pattison P (1996) Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p^*. Psychometrika 61:401–425
56. Wasserman S, Robins G (2005) An Introduction to Random Graphs, Dependence Graphs, and p*. In: Carrington P, Scott J, Wasserman S (eds) Models and Methods in Social Network Analysis. Cambridge University Press, New York, pp 148–161
57. Zhao Y, Robins G (2006) Multiple networks: Comparing QAP and exponential random graph (p*) models. Sunbelt XXVI International Social Networks Conference, Vancouver, April 2006

Books and Reviews

Agneessens F, Roose H, Waege H (2004) Choices of theatre events: p^* models for affiliation networks with attributes. Metodoloski Zvezki 1:419–439

Anderson C, Wasserman S, Crouch B (1999) A p^* primer: Logit models for social networks. Soc Netw 21:37–66

Baddeley A, Möller J (1989) Nearest-neighbour Markov point processes and random sets. Int Stat Rev 57:89–121

Boer P, Huisman M, Snijders TAB, Zeggelink E (2003) StOCNET: an open software system for the advanced analysis of social networks. ProGAMMA/ICS, Groningen

Brieger R (1981) Comment on An exponential family of probability distributions for directed graphs. J Am Stat Assoc 76:51–53

Butts CT (2008) Social network analysis: A methodological introduction. Asian J Soc Psychol 11:13–41

Contractor N, Wasserman S, Faust K (2006) Testing multi-theoretical multilevel hypotheses about organizational networks: An analytic framework and empirical example. Acad Manag J 31:681–703

Frank O (1981) A survey of statistical methods for graph analysis. Sociol Methodol 11:110–155

Frank O (1991) Statistical analysis of change in networks. Stat Neerlandica 95:283–293

Frank O, Nowicki K (1993) Exploratory statistical analysis of networks. Quo Vadis, Graph Theory? Ann Discret Math 55:349–366

Handcock M, Hunter D, Butts C, Goodreau S, Morris M (2006) Statnet: An R package for the statistical analysis and simulation of social networks. University of Washington, Washington

Koehly L, Goodreau S, Morris M (2004) The link between exponential random graph models and loglinear models for networks. Sociol Methodol 34:241–270

Lazega E, van Duijn M (1997) Position in formal structure, personal characteristics and choices of advisors in a law firm: A logistic regression model for dyadic network data. Soc Netw 19:375–397

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chkovskii D, Alon U (2002) Network motifs: simple building blocks of complex networks. Science 298:824–827

Monge P, Contractor N (2003) Theories of communication networks. Oxford University Press, New York

Newman M (2003) The structure and function of complex networks. SIAM Rev 45:167–256

Pattison P, Robins GL (2008) Probabilistic network theory. In: Rudas T (ed) Handbook of Probability: Theory with Applications. Sage, Thousand Oaks, pp 291–312

Pattison P, Wasserman S (2002) Multivariate random graph distributions: applications to social network analysis. In: Hagberg J (ed) Contributions to social network analysis, information theory and other topics in statistics: A festschrift in honour of Ove Frank. Department of Statistics, University of Stockholm, Stockholm, pp 74–100

Robins G, Pattison P (2001) Random graph models for temporal processes in social networks. J Math Sociol 25:5–41

Robins G, Pattison P, Woolcock J (2004) Models for social networks with missing data. Soc Netw 26:257–283

Schweinberger M, Snijders TAB (2003) Settings in social networks: A measurement model. Sociol Methodol 33:307–341

Wang P, Robins G, Pattison P (2006) PNet: Program for the estimation and simulation of *p**exponential random graph models, User Manual. School of Behavioural Science, University of Melbourne, Melbourne

Social Networks and Granular Computing

CHURN-JUNG LIAU
Institute of Information Science, Academia Sinica,
Taipei, Taiwan

Article Outline

Glossary

Granulation The process of drawing a set of objects (or points) together by indiscernibility, similarity, proximity, or functionality.

Functional granulation The granulation process is functional if it is based on the attributes of the objects. It is called functional granulation because each attribute is a function from the set of objects to the set of values.

Relational granulation The granulation process is relational if it is based on the relationships between objects.

Rough set A rough set is defined by the lower and upper approximations of a concept. The lower approximation contains all elements that necessarily belong to the concept, while the upper approximation contains those that possibly belong to the concept. In rough set theory, a concept is considered a classical set.

Social network A social network is comprised of a set of actors, called the domain, and a family of relations on the domain. It is usually represented as a graph, where each node represents an actor and an edge between two nodes represents a relational tie between these two actors. An edge can be labeled with the relation it represents.

Positional equivalence Two actors are in equivalent positions if their "pattern" of relationships with other actors is the same.

Structural equivalence Two actors are structurally equivalent if they are related to the same actors.

Regular equivalence Two actors are regularly equivalent if they are equally related to equivalent actors.

Exact equivalence Two actors are exactly equivalent if they are related to the same number of equivalent actors.

Modal logic Modal logic was originally developed as a type of philosophical logic for reasoning about necessity and possibility. However, it has been extended to broadly cover a family of logics for reasoning about modalities including tense, obligation, belief, and knowledge. Semantically, it is also a powerful mathematical discipline that deals with (restricted) description languages for discussing various kinds of relational structures, where a relational structure comprises a set of elements and a collection of relations on that set.

Hybrid logic Hybrid logic is a branch of modal logic that allows direct reference to the elements in a relational structure. Traditionally, only the properties of the elements could be represented by modal logic formulas.

Definition of the Subject

Granular computing (GrC) is a problem-solving concept deeply rooted in human thinking. Hence, it has played a major role in solving many important problems throughout the history of mathematics. GrC is concerned with the processing of information granules, which are groups of objects drawn together by indiscernibility, similarity, proximity, or functionality [37]. The process of forming information granules is called granulation. If the process is based totally on the attributes of the objects, it is called *functional granulation*, since attributes are mathematical functions from the set of objects to the set of values; if, in addition, the granulation process is also based on the relationship between objects, it is called *relational granulation* [11,20].

Interestingly, social scientists have applied the techniques of relational granulation (albeit by different names) to positional analysis in social networks [7,8,10,19,35]. Social network analysis (SNA) is a methodology used extensively in social and behavioral sciences, as well as in political science, economics, organization theory, and industrial engineering [15,31,34]. Positional analysis of a social network tries to find similarities between actors in the network. While many traditional clustering methods are based on the attributes of the individual actors, SNA is more concerned with the structural similarity between the actors. In SNA, a category, called a *social role* or *social position*, is defined in terms of the similarities of the patterns of relations among the actors, rather than the attributes of the actors.

Analyzing a social position from a GrC perspective has two major advantages. On the one hand, it extends the application scope of the GrC methodology. On the other hand, we can logically transform relational granulation of social networks into functional granulation; as a result, the well-developed techniques of data table analysis can be applied to social position analysis. Thus, this is an important cross-fertilization of GrC and SNA.

Introduction

GrC encompasses a wide range of techniques in computational intelligence, such as interval computing, classification, cluster analysis, and fuzzy and rough set theories. Among them, rough set theory [26,27] is one of the most well-known methods with successful applications in data analysis and uncertainty management. The theory was first proposed by Z. Pawlak as a mathematical tool to deal with vague concepts [26]. The basic assumption of the theory is that the effective use of knowledge is based on the capability to classify objects. Thus, knowledge consists of a family of classifications of a domain of interest. A classification of a domain is represented as a partition of (or an equivalence relation on) the domain. Each equivalence class of a classification is the elementary knowledge about the classification. For example, if the domain is classified according to colors, then the elementary knowledge may include "red", "green", "blue", etc. In this sense, a classification is derivable from a set of features or attributes of objects. An important task of data analysis is to define a concept via a set of features. If the concept can be precisely defined as the objects with some specific attribute values, then it is an exact concept with respect to the given attributes. However, in practice, most concepts are vague and can not be defined precisely by a given set of attributes. In such cases, rough set theory provides an effective way to approximate the vague concepts with some exact concepts in the domain. Because of its applications in data mining, rough set theory has flourished since 1990.

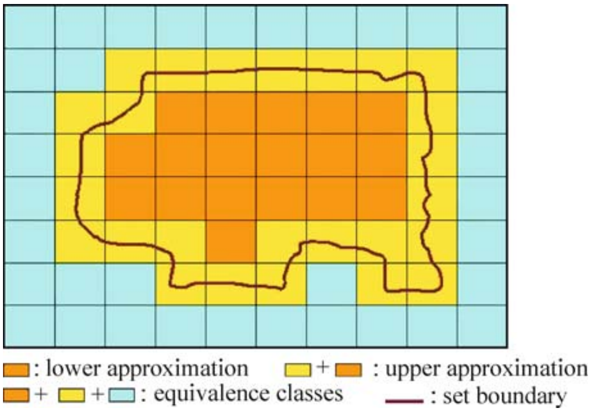
At about the time rough set theory was proposed, the pioneer of fuzzy set theory, L.A. Zadeh, emphasized the importance of information granularity in fuzzy reasoning [36]. Subsequently, T.Y. Lin proposed the term "GrC" to unify different techniques developed for processing information granules. From the GrC perspective, both classification in rough set theory and fuzzification in fuzzy set theory are types of granulation.

In rough set theory, objects are partitioned into equivalence classes based on their attribute values. While such values essentially represent functional information associated with the objects, more complicated information can be utilized in the classification of objects. In particular, relational information between objects can play an important role in the granulation process. The information is defined by general binary relations, which are extensions of the functional attributes of the objects. Geometrically, such granulation is derived from the neighborhood system of topological spaces [32], where each point/object is assigned at most one neighborhood/granule. This kind of granulation is called *relational granulation*, while granulation based on attribute values only is called *functional granulation* [11,20].

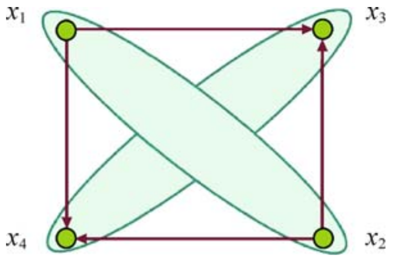
Relational granulation is especially useful for the analysis of network data because a network is a collection of relations between nodes. A concrete instance of the application of relational granulation is positional analysis of social networks. Positional analysis attempts to find actors occupying the same position in a social network based on the pattern of their relationships to other actors. Depending on the different patterns of relationships, different notions of positional equivalence have been proposed [19]. These notions can be easily regarded as instances of relational granulation. In this article, we consider three of the most important notions of positional equivalence – *structural equivalence*, *regular equivalence*, and *exact equivalence*.

Recently, it was shown that social positions based on regular equivalence can be syntactically expressed as well-formed formulas (wff) in a kind of modal logic [22]. Thus, actors occupying the same social position based on regular equivalence will satisfy the same set of modal formulas. Traditionally, modal logic has been considered the logic for reasoning about modalities, such as necessity, possibility, time, action, belief, knowledge, and obligation. However, semantically, it is essentially a language for describing relational structures [5]. A relational structure is simply a set combined with a collection of relations on that set. Thus, social networks are mathematically equivalent to relational structures. The logical characterization of social positions implies that relational granulation for positional analysis can be transformed into a functional granulation process. Indeed, since each actor in a social network may satisfy or falsify a modal formula, a modal formula can be seen as a binary attribute associated with each actor. Therefore, based on the results in [22], two actors are regularly equivalent if they have the same values with respect to all attributes defined by the language of the particular modal logic.

By extending previous results, we can find logical characterizations of different positional equivalences and transform them into a functional granulation process. Consequently, the techniques of rough set-based data analysis can also be applied to positional analysis. However, since the set of wffs of a modal logic is usually infinite, the functional granulations corresponding to positional equivalences need to consider infinite sets. To circumvent this problem, we also propose a new definition of positional equivalence, called *observational equivalence*. Let Σ be a set of wffs in a logical language that denotes the set of observable properties of the actors in a network. Two actors a and b are observationally equivalent with respect to Σ, if for each wff $\varphi \in \Sigma$, a satisfies φ iff b satisfies φ.

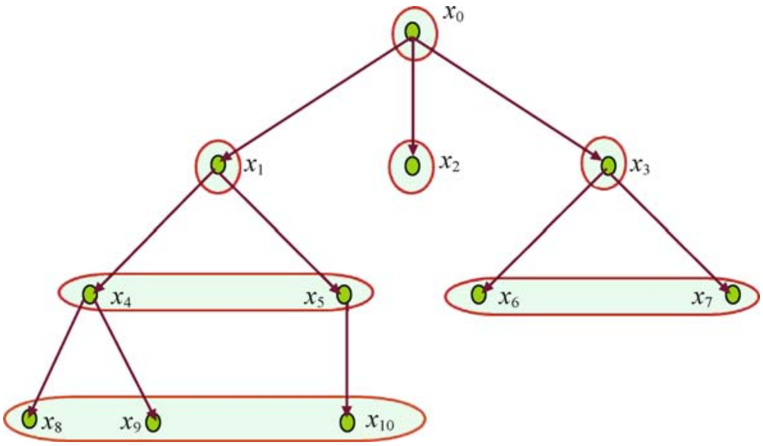


Social Networks and Granular Computing, Figure 1
A rough set: the set delimited by the *curve* (the set boundary) can not be represented exactly as a union of the basic building blocks. The *blocks* totally included in the set form the lower approximation of the set, while the *blocks* that intersect with the set (including the lower approximation) form the upper approximation of the set

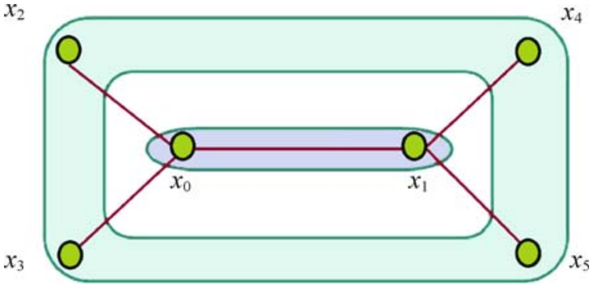


Social Networks and Granular Computing, Figure 2
Structural equivalence: the social network contains 4 actors denoted by the *nodes* and a single binary relation denoted by the *arcs*. Both the out-neighborhoods of x_1 and x_2 are $\{x_3, x_4\}$, and their common in-neighborhood is the empty set. On the other hand, both the in-neighborhoods of x_3 and x_4 are $\{x_1, x_2\}$, and their common out-neighborhood is the empty set. Thus, $x_1 \cong_s x_2$ and $x_3 \cong_s x_4$. The *green shadowed areas* indicate the equivalence classes

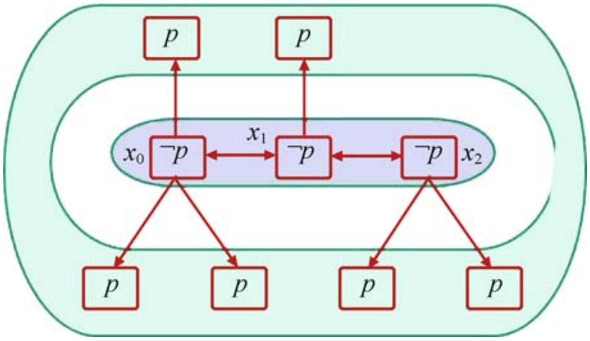
The remainder of this article is organized as follows. In Sect.“ Mathematical Preliminaries”, we introduce the background knowledge and mathematical notations of sets and relations. In Sects. “Rough Set Theory – Functional Granulation” to “Modal Logic – The Bridge”, we review rough set theory, social positional analysis, and modal logic respectively. In Sect. “From Relational Granulation to Functional Granulation”, we present the logical characterizations of three main positional equivalences and the definition of observational equivalence. Finally, in Sect. “Conclusion and Future Directions”, we present our conclusions and indicate some future research directions.



Social Networks and Granular Computing, Figure 3
Regular equivalence in a tree-structured social network: the root x_0 is discernible from other nodes because its in-neighborhood is empty, whereas all the other nodes have nonempty in-neighborhoods. Nodes x_1, x_2, x_3 are discernible from other nodes because they are the only nodes connected to x_0. Moreover, they are mutually discernible, since x_2 has no children, x_1 has children and grand-children, and x_3 has children, but not grand-children. Note that x_4 and x_5 are indiscernible, since they have the same in-neighborhood and equivalent out-neighborhoods; x_6 and x_7 are indiscernible, since they have the same in-neighborhood and both have empty out-neighborhoods; and x_8, x_9, and x_{10} are indiscernible, since they have equivalent in-neighborhoods and all have empty out-neighborhoods



Social Networks and Granular Computing, Figure 4
Exact equivalence: the social network is a undirected graph. In other words, the binary relation is symmetry. The exact equivalence partitions the nodes into two classes, $C_0 = \{x_0, x_1\}$ and $C_1 = \{x_2, x_3, x_4, x_5\}$. To verify that it is indeed an exact equivalence, we can check that both x_0 and x_1 connect to one C_0 node and two C_1 nodes, and all of x_2, x_3, x_4, and x_5 connect to one C_0 node



Social Networks and Granular Computing, Figure 5
A social network represented as a Kripke model: we assume the underlying modal language contains one propositional symbol p and one relational symbol r. Since $\{x_0, x_1, x_2\}$ is an equivalence class in the shown regular equivalence, the three nodes all satisfy the same set of PMML wffs. For example, they all satisfy $\neg p \wedge \langle r \rangle p$. However, they are not exactly equivalent, so they are distinguishable by GML wffs. For example, x_0 satisfies both $\langle r \rangle_1 p$ and $\langle r \rangle_2 p$, x_1 does not satisfy $\langle r \rangle_1 p$ or $\langle r \rangle_2 p$, and x_2 satisfies $\langle r \rangle_1 p$ but not $\langle r \rangle_2 p$

Mathematical Preliminaries

In this section, we introduce the basic knowledge about sets and relations used in this article.

Sets

A set is a collection of objects, called the elements of the set. We use capitals A, B, U, X, etc. to denote sets. The elements of a set are denoted by lower-case letters and we write $x \in U$ if x is an element of the set U. We usually use two notations to denote a set. The first one lists its elements. For example, we can write $\{x_1, x_2, \ldots, x_n\}$ for a finite set, where n is a positive integer, and $\{x_1, x_2, \ldots\}$ for an infinite set. The second notation specifies the defining property of the set. For example, we can write $\{x \mid x = 2n, n \in Z\}$ for the set of even numbers. The number of elements in a set is called its cardinality. The cardinality of U is denoted by $|U|$. A set U is called an empty set if $|U| = 0$, and a singleton if $|U| = 1$. We denote the empty set by $\emptyset$.

The basic operations on sets are defined as follows:

1. Intersection of X and Y: $X \cap Y = \{u \mid u \in X \text{ and } u \in Y\}$,
2. Union of X and Y: $X \cup Y = \{u \mid u \in X \text{ or } u \in Y\}$,
3. Cartesian product of X and Y: $X \times Y = \{(x, y) \mid x \in X \text{ and } y \in Y\}$, and
4. Difference of X by Y: $X - Y = \{u \mid u \in X \text{ and } u \notin Y\}$.

Since intersection, union, and the Cartesian product are associative (i. e., $(X \cap Y) \cap Z = X \cap (Y \cap Z)$ etc.), we can eliminate the parentheses and apply the operations to more than two sets. For example, the Cartesian product of the sets $U_1, U_2, \ldots, U_k$ is denoted by

$$U_1 \times U_2 \times \cdots \times U_k = \{(x_1, x_2, \ldots, x_k) \mid x_i \in U_i, \forall 1 \leq i \leq k\} ,$$

where each element is called a k-tuple. A 2-tuple is also called a pair. If $U_1 = U_2 = \cdots = U_k = U$, the Cartesian product is written as U^k. A set X is said to be a subset of another set Y, denoted by $X \subseteq Y$, if for all $x \in X$, it is also the case that $x \in Y$. A function f from the set X to the set Y, denoted by $f : X \to Y$, is a mapping that associates each element in X with an element in Y. The set X is called the domain of the function, and Y is called the range of f. For each $x \in X$, we write $f(x)$ for the element in Y that is associated with x by f and call it the image of x under f. Also, the image of X under f is defined as $f(X) = \{f(x) \mid x \in X\}$.

Multisets

For sets, repeat occurrences of the same elements are only counted once. Thus, the set $\{x, x\}$ is the same as the set $\{x\}$. Sometimes, the number of occurrences of an element is important. In such cases, we consider multisets (or bags). A multiset is formally defined as a pair $M = (U, m)$, where U is some set and $m : U \to N$ is a function from U to the set $N = \{0, 1, 2, 3, \ldots\}$ of natural numbers. U is called the universe and for each element $x \in U$, $m(x)$ is the multiplicity (that is, the number of occurrences) of x in M. A set can be considered as a special multiset in which $m(x) = 0$ or 1 for each x in the universe. By generalizing the notation of set membership, we can write $x \in M$ if $m(x) > 0$. The cardinality of M is defined as $|M| = \sum_{x \in U} m(x)$. Let $M_1 = (U, m_1)$ and $M_2 = (U, m_2)$ be two multisets in the same universe; then, $M_1 \subseteq M_2$ iff $m_1(x) \leq m_2(x)$ for all $x \in U$. The intersection, union, and difference between $M_1 = (U, m_1)$ and $M_2 = (U, m_2)$ are defined as follows:

1. $M_1 \cap M_2 = (U, m)$, where $m(x) = \min(m_1(x), m_2(x))$ for all $x \in U$,
2. $M_1 \cup M_2 = (U, m)$, where $m(x) = \max(m_1(x), m_2(x))$ for all $x \in U$,
3. $M_1 - M_2 = (U, m)$, where $m(x) = (m_1(x) - m_2(x)) \cdot \chi(m_1(x) - m_2(x))$ for all $x \in U$, and

$$\chi(k) = \begin{cases} 1, & \text{if } \quad k > 0, \\ 0, & \text{if } \quad k \leq 0 . \end{cases}$$

The Cartesian product of $M_1 = (U_1, m_1)$ and $M_2 = (U_2, m_2)$ is defined as $M_1 \times M_2 = (U_1 \times U_2, m)$, where $m(x, y) = m_1(x) m_2(y)$ for all $x \in U_1$ and $y \in U_2$.

Relations

In mathematics, a subset of $U_1 \times U_2 \times \cdots \times U_k$ is called a k-ary relation among the sets $U_1, U_2, \ldots, U_k$, and a subset of U^k is called a k-ary relation on U. We are particularly interested in binary relations. According to the definition, a binary relation on U is a set of pairs of elements in U. A pair (x, y) in a binary relation means that x is related to y by the relation. We use lower-case Greek letters, α, β, ρ, etc. to denote relations. The basic operations on binary relations include the three aforementioned set-theoretic operations – intersection, union, and difference, as well as converse and composition. The converse of a binary relation $\alpha \subseteq U_1 \times U_2$ is $\alpha^- \subseteq U_2 \times U_1$ such that

$$\alpha^- = \{(x, y) \mid (y, x) \in \alpha\} .$$

The composition of two binary relations $\alpha \subseteq U_1 \times U_2$ and $\beta \subseteq U_2 \times U_3$ is $\alpha\beta \subseteq U_1 \times U_3$ such that

$$\alpha\beta = \{(x, y) \mid \exists z \in U, (x, z) \in \alpha \wedge (z, y) \in \beta\} .$$

A binary relation $\rho \subseteq U^2$ is said to be reflexive if $(x, x) \in \rho$ for all $x \in U$; ρ is symmetric if for all $x, y \in U$, $(x, y) \in \rho$ implies $(y, x) \in \rho$; ρ is anti-symmetric if for all $x, y \in U$, $(x, y) \in \rho$ and $(y, x) \in \rho$ implies $x = y$; and ρ is transitive if for all $x, y, z \in U$, $(x, y) \in \rho$ and $(y, z) \in \rho$ implies $(x, z) \in \rho$. A binary relation that is reflexive, antisymmetric, and transitive is called a partial order. Let $\leq$ be a partial order on the universe U and $X \subseteq U$, then, $u \in U$ is called an upper bound (resp. lower bound) of X if for all $x \in X$, $x \leq u$ (resp. $u \leq x$). The least upper bound (resp. the greatest lower bound) of X is the upper bound (resp. lower bound) u of X such that for any upper bounds (resp. lower bounds) u' of X, $u \leq u'$ (resp. $u' \leq u$). The least upper bound (resp. the greatest lower bound) of X is called its supremum (resp. infimum), and is denoted by $\sup(X)$ (resp. $\inf(X)$). For any X (even though X is finite), $\sup(X)$ and $\inf(X)$ do not necessarily exist. If for all $x, y \in U$, $\sup(\{x, y\})$ and $\inf(\{x, y\})$ exist, then $\sup(\{x, y\})$ is called the join of x and y and denoted by $x \sqcup y$, while $\inf(\{x, y\})$ is called the meet of x and y and denoted by $x \sqcap y$. The structure $(U, \sqcap, \sqcup)$ is called a lattice. A subset $X \subseteq U$ is a sublattice of U if for any $x, y \in X$, $x \sqcap y$ and $x \sqcup y \in X$.

A binary relation that is reflexive, symmetric, and transitive is called an equivalence relation, which yields a partition of the universe. A partition of U is a class of subsets of U, $\{X_i \mid i \in I\}$ for some index set I, where $\bigcup_{i \in I} X_i = U$ and $X_i \cap X_j = \emptyset$ for any $i \neq j$. Let ρ be an equivalence relation on U, and let $x \in U$; then, the equivalence class containing x is defined as

$$[x]_\rho = \{y \in U \mid (x, y) \in \rho\} .$$

Obviously, $\{[x]_\rho \mid x \in U\}$ forms a partition of the universe. For any $X \subseteq U$, we denote $[X]_\rho$ by the set $\{[x]_\rho \mid x \in X\}$ and $[\![X]\!]_\rho$ by the multiset $([U]_\rho, m)$ such that $m([x]_\rho) = |[x]_\rho \cap X|$ for all $x \in U$.

Equivalence relations on a fixed domain are partially ordered by set-inclusion. Let ρ_1 and ρ_2 be two equivalence relations and define $\rho_1 \leq \rho_2$ iff $\rho_1 \subseteq \rho_2$; then, $\leq$ is a partial order on the set of all equivalence relations on a fixed domain. We say that ρ_1 is finer than ρ_2 or ρ_2 is coarser than ρ_1 if $\rho_1 \leq \rho_2$. Given a binary relation α, the composition of α with itself k times is denoted by α^k and the transitive closure of α is defined as $\alpha^\infty = \bigcup_{k \geq 1} \alpha^k$. Let ρ_1 and ρ_2 be two equivalence relations; then it can be shown that $\rho_1 \cap \rho_2$ and $(\rho_1 \cup \rho_2)^\infty$ are, respectively, the meet and the join of ρ_1 and ρ_2.

Rough Set Theory – Functional Granulation

The basic construct of rough set theory is the approximation space, which is a pair (U, ρ), where U is a finite set of objects (the universe) and ρ is an equivalence relation on U. From the GrC perspective, the equivalence classes of ρ are information granules of the approximation space.

In rough set theory, a concept is represented as a subset of the universe. For any concept $X \subseteq U$, we can associate the following two subsets with X,

$$\underline{\rho}X = \{x \in U \mid [x]_\rho \subseteq X\}$$
$$\overline{\rho}X = \{x \in U \mid [x]_\rho \cap X \neq \emptyset\} ,$$

where $\underline{\rho}X$ and $\overline{\rho}X$ are called the ρ-lower and ρ-upper approximation of X, respectively. From a practical viewpoint, ρ can be considered as an indiscernibility relation; thus, for a given concept X, we can only know that X contains all the elements in $\underline{\rho}X$ and does not contain any element outside $\overline{\rho}X$. In other words, the concept X cannot be defined exactly in the approximation space. The pair $(\underline{\rho}X, \overline{\rho}X)$ is considered a rough approximation of X, and any such pair is called a rough set (see Fig. 1). A concept X is called an exact concept in the approximation space if $\underline{\rho}X = X = \overline{\rho}X$; otherwise, it is called a rough concept.

Rough set theory is very useful in the analysis of data tables. A data table is a pair $S = (U, F)$, where U is a nonempty, finite set (the universe) and F is a nonempty, finite set of primitive attributes. Every $f \in F$ is a total function $f \colon U \to V_f$, where V_f denotes the possible values of f. If $E = \{f_1, \dots, f_n\} \subseteq F$, then $V_E = V_{f_1} \times \cdots \times V_{f_n}$. Thus, E is also considered a function from U to V_E. An equivalence relation $\mathrm{IND}(E)$ is associated with every subset of attributes $E \subseteq F$, and defined by

$$(x, y) \in \mathrm{IND}(E) \Leftrightarrow f(x) = f(y) \quad \forall f \in E .$$

$\mathrm{IND}(E)$ is called an indiscernibility relation. We write $\mathrm{IND}(f)$ instead of $\mathrm{IND}(\{f\})$ for all $f \in F$. It is easy to show that $\mathrm{IND}(E) = \bigcap_{f \in E} \mathrm{IND}(f)$. Since $\mathrm{IND}(E)$ is an equivalence relation, $(U, \mathrm{IND}(E))$ is an approximation space and we can define the $\mathrm{IND}(E)$-lower and $\mathrm{IND}(E)$-upper approximation of X for any $X \subseteq U$. We write $\underline{E}X$ and $\overline{E}X$ instead of $\underline{\mathrm{IND}(E)}X$ or $\overline{\mathrm{IND}(E)}X$. Obviously, exact concepts in $(U, \mathrm{IND}(E))$ are those definable using only attributes in E. In effect, the indiscernibility relation $\mathrm{IND}(E)$ granulates (or partitions) the universe according to the functions in E; thus, it results in a functional granulation of the universe.

The definitions are used in the analysis of dependency between attributes in a data table. Let us say that attribute E_2 depends on E_1, denoted by $E_1 \Rightarrow E_2$, iff $\mathrm{IND}(E_1) \subseteq \mathrm{IND}(E_2)$, i. e., any two objects in U with the same attribute values in E_1 will also have the same attribute values in E_2. It is easy to show that $E_1 \Rightarrow E_2$ iff $\underline{E_1}X = X$ for all X that are equivalence classes of $\mathrm{IND}(E_2)$. Given an attribute $f \in F$, if $F - \{f\} \Rightarrow \{f\}$, then f can be eliminated from F without influencing the definability of F. In other words, by using the set of attributes, f is redundant for classification purposes. By repeated elimination of redundant attributes, we can eventually obtain an irreducible set of attributes with the same classification capability as the original set of attributes. Such an irreducible set of attributes is called a reduct in rough set theory. Note that the definition of dependency is relative to a data table, so it does not necessarily mean that an intrinsic connection exists between the inter-dependent attributes.

Positional Analysis – Relational Granulation

Social networks are defined by actors and relations (or nodes and edges in terms of graph theory) [15]. A social network is generally defined as a relational structure $\mathfrak{N} = (A, (\alpha_i)_{i \in I})$, where A is the set of actors in the network, I is an index set, and for each $i \in I$, $\alpha_i \subseteq A^{k_i}$ is a k_i-ary relation on the domain A. If $k_i = 1$, then α_i is also called an attribute. In practice, most SNA literature considers a simplified version of social networks with only binary relations. For ease of presentation, we focus on a social network with only unary and/or binary relations. Thus, the social network considered in this article is a structure $\mathfrak{N} = (A, (P_i)_{i \in I}, (\alpha_j)_{j \in J})$, where A is a *finite set* of actors, $P_i \subseteq A$ for all $i \in I$, and $\alpha_j \subseteq A \times A$ for all $j \in J$. In terms of graph theory, $\mathfrak{N}$ is a labeled graph, where A is a set of nodes labeled with subsets of I, and each α_j denotes a set of (labeled) edges. For each $a \in A$, the out-neighborhood and in-neighborhood of a with respect to a binary relation α, denoted respectively by $N^+_\alpha(a)$ and $N^-_\alpha(a)$, are defined as follows:

$$N^+_\alpha(a) = \{b \in A \mid (a, b) \in \alpha\} ,$$
$$N^-_\alpha(a) = \{b \in A \mid (b, a) \in \alpha\} .$$

If ρ is an equivalence relation on A and a is an actor, the ρ-equivalence class of a is equal to its neighborhood, i. e., $[a]_\rho = N^+_\rho(a) = N^-_\rho(a)$. Note that the latter equality holds because of the symmetry of ρ.

Several equivalence relations have been proposed for exploring the similarity between actors' roles. The simplest definition of positional equivalence is the concept of structural equivalence proposed in [21], which states that two actors are positionally equivalent if they are related to the same individuals (see Fig. 2).

Definition 1 Let $\mathfrak{N} = (A, (P_i)_{i \in I}, (\alpha_j)_{j \in J})$ be a social network and ρ be an equivalence relation on A; then ρ is a (strong) structural equivalence with respect to $\mathfrak{N}$ if $(a, b) \in \rho$ implies

1. $a \in P_i$ iff $b \in P_i$ for all $i \in I$, and
2. $N^+_{\alpha_j}(a) = N^+_{\alpha_j}(b)$ and $N^-_{\alpha_j}(a) = N^-_{\alpha_j}(b)$ for all $j \in J$.

By the definition, there may exist more than one structural equivalence for a given network. However, it has been shown that there always exists a maximum (i. e., coarsest) structural equivalence for a network [19]. In fact, the set of all structural equivalences form a sublattice of all equivalence relations on A. Indeed, if ρ_1 and ρ_2 are structural equivalences with respect to $\mathfrak{N}$, so too are $\rho_1 \sqcap \rho_2$ and $\rho_1 \sqcup \rho_2$. Since the set of all equivalence relations on a finite domain is finite, the join of all structural equivalences is the coarsest structural equivalence with respect to a network. Let $\mathfrak{N} = (A, (P_i)_{i \in I}, (\alpha_j)_{j \in J})$ be a social network and ρ be the coarsest structural equivalence with respect to $\mathfrak{N}$; then, two actors $x, y \in A$ are said to be structurally equivalent, denoted by $x \cong_s y$, if $(x, y) \in \rho$.

Although structural equivalence is conceptually simple, it is sometimes restrictive. Regular equivalence relaxes the requirement that equivalent actors must be connected with identical actors, and suggests that actors occupy the same position if they are connected to positionally equivalent actors. Regular equivalence has been studied extensively [7,8,10,19,35], and there are several alternative definitions of the concept. We present two here. The first is based on the characterization given by Boyd and Everett [8], which states that an equivalence relation ρ is a *regular equivalence* with respect to a binary relation α if it commutes with α, i. e.

$$\alpha\rho = \rho\alpha .$$

By this definition, if ρ is a regular equivalence with respect to α and $(a, b) \in \rho$, then for each $c \in N^+_\alpha(a)(\text{resp.} N^-_\alpha(a))$, there exists $c' \in N^+_\alpha(b)(\text{resp.} N^-_\alpha(b))$ such that $(c, c') \in \rho$. The property naturally leads to an alternative definition of regular equivalence [19], which states that an equivalence relation ρ is a *regular equivalence* with respect to a binary relation α if for $a, b \in A$,

$$(a, b) \in \rho \Rightarrow [N^+_\alpha(a)]_\rho = [N^+_\alpha(b)]_\rho \quad \text{and} \quad [N^-_\alpha(a)]_\rho = [N^-_\alpha(b)]_\rho .$$

According to this definition, if a and b are regularly equivalent, then they are connected to equivalent neighborhoods (see Fig. 3). Obviously, the above definitions are equivalent. Thus, we have the following definition.

Definition 2 Let $\mathfrak{N} = (A, (P_i)_{i \in I}, (\alpha_j)_{j \in J})$ be a social network and ρ be an equivalence relation on A; then ρ is a regular equivalence with respect to $\mathfrak{N}$ if

1. $(a, b) \in \rho$ implies $a \in P_i$ iff $b \in P_i$ for all $i \in I$; and
2. ρ is a regular equivalence with respect to α_j for all $j \in J$.

Analogous to the case of structural equivalences, the join of two regular equivalences is still a regular equivalence, even though their meet is not necessarily a regular equivalence. Consequently, we can find the coarsest regular equivalence of a network. Then, two actors, x and y, are regularly equivalent, denoted by $x \cong_r y$, if (x, y) is in the coarsest regular equivalence of the network.

For regular equivalence, only the occurrence or non-occurrence of a position in the neighborhood of an actor matters. However, the number of occurrences is sometimes an important factor in positional analysis. In such cases, a number restriction can be added to the definition of regular equivalences. An equivalence relation ρ is an *exact equivalence* with respect to a binary relation α if for $a, b \in A$,

$$(a, b) \in \rho \Rightarrow [\![N_\alpha^+(a)]\!]_\rho = [\![N_\alpha^+(b)]\!]_\rho \quad \text{and} \quad [\![N_\alpha^-(a)]\!]_\rho = [\![N_\alpha^-(b)]\!]_\rho \,.$$

By replacing the set equality in the definition with the multiset equality, the number of equivalent neighbors must be the same for two actors to be considered position-equivalent (see Fig. 4).

Definition 3 Let $\mathfrak{N} = (A, (P_i)_{i \in I}, (\alpha_j)_{j \in J})$ be a social network and ρ be an equivalence relation on A; then, ρ is an exact equivalence with respect to $\mathfrak{N}$ if

1. $(a, b) \in \rho$ implies $a \in P_i$ iff $b \in P_i$ for all $i \in I$; and
2. ρ is an exact equivalence with respect to α_j for all $j \in J$.

The set of all exact equivalences also forms a lattice, but it is not a sublattice of the set of all equivalence relations [10]. Consequently, the coarsest exact equivalence for a network exists and we can define two actors x and y as exactly equivalent, denoted by $x \cong_e y$, if they belong to the coarsest exact equivalence. The partition produced by an exact equivalence is called an equitable partition or a divisor of a graph [19].

From the above definition, it is clear that all the attributes in a social network are binary. In other words, for each attribute P_i, an actor either satisfies P_i (has the value 1) or falsifies P_i (has the value 0). When a social network $\mathfrak{N} = (A, (P_i)_{i \in I}, (\alpha_j)_{j \in J})$ has the property $J = \emptyset$, it is in fact a kind of data table. In this case, all three kinds of positional equivalence are reduced to the indiscernibility relation. This means that, if information about the actors' relationship is not available, social positions are fully determined by the attributes of the actors. Thus, without relational information, functional granulation is sufficient for positional analysis. However, if relational information is available, two actors can be further differentiated by their relationship with other actors, even though all their attributes have the same values. In this sense, the definition of positional equivalence is based on the process of relational granulation.

Modal Logic – The Bridge

Modal logics were originally developed as formalizations for reasoning about modalities. The formal study started with alethic modal logics (the logics of necessity and possibility). The modalities □ (for necessity) and ◇ (for possibility) have now become standard notations in modal logic literature, and many variants of modal logic have been proposed to deal with different classes of modalities; for example, temporal logic for tense operators, deontic logic for obligation and permission, dynamic logic for actions, and epistemic logic for beliefs and knowledge. The development of these logics was motivated by philosophical enquiry as well as by technical applications in computer science, artificial intelligence, and economic game theory.

While modal logics were initially presented in the form of reasoning systems, the invention of relational semantics has been influential in the continuing development of the field [17,18]. The semantics, proposed independently by J. Hintikka, S. Kanger, and S. Kripke, and now known as Kripke semantics, show that modal logics are in fact logics for reasoning about relational structures. From this semantic perspective, standard modal logics can be regarded as fragments of first-or second-order predicate logics, whereby the necessity and possibility modalities correspond to universal and existential quantifiers, respectively. In spite of this correspondence, quantification in modal logic tends to be bounded in some way to worlds that are "relevant to" or "accessible from" the current one. Consequently, although a number of properties of modal logics follow immediately from those of their classical quantificational counterparts, the modal operators typically have less expressive power than full quantification. This results in many interesting properties not available in classical predicate logic. One of the most striking results is that semantic invariances between models are actually various forms of *bisimulation*, which preserve the local properties of worlds and their transition patterns [5]. Interestingly, it has been shown that bisimulation in Kripke models corresponds exactly to regular equivalence in social networks [22]. The implication of the results is that position-equivalent actors can be characterized by an appropriate set of modal formulas. As each modal formula can be regarded as a binary attribute, from the GrC perspective, such characterization transforms relational granulation into functional granulation.

In this section, we present three modal logics – a multimodal logic, a graded modal logic, and a hybrid logic. The presentation of a modal logic, as with any other formal logic, requires the specification of its syntax and semantics. The syntactic aspect includes the language of the logic – its alphabet and formation rules for wffs, as well as a deductive system for logical reasoning. For the purpose of this article, we need only present the language part. For the semantic aspect, we have to define the models within which the wffs can be interpreted, as well as the satisfaction of a wff in a model.

Propositional Multi-Modal Logic

We start with propositional modal logic (PML) [9]. The alphabet of PML consists of a set of propositional symbols, PV, and the logical symbols $\neg$(negation), $\wedge$(and), $\vee$(or), $\supset$(material implication), and $\Diamond$(possibility); $\neg$, $\wedge$, $\vee$, and $\supset$ are called Boolean connectives, whereas $\Diamond$ is the only primitive modality of PML. The set of wffs of PML is the smallest set containing PV that satisfies the following conditions:

- if φ is a wff, then $\neg\varphi$ and $\Diamond\varphi$ are wffs,
- if φ and ψ are wffs, then $\varphi \wedge \psi$, $\varphi \vee \psi$, and $\varphi \supset \psi$ are wffs.

As usual, we abbreviate $(\varphi \supset \psi) \wedge (\psi \supset \varphi)$ as $\varphi \equiv \psi$; $\neg \Diamond \neg\varphi$ as $\Box\varphi$; any tautology $p \vee \neg p$ as $\top$; and any contradiction $p \wedge \neg p$ as $\bot$. A Kripke model for PML is a triple $\mathfrak{M} = (W, \alpha, V)$, where W is a set of possible worlds, α is a binary relation on W, called an accessibility relation, and $V\colon W \times PV \to \{0, 1\}$ is a truth assignment for evaluating the truth value of each propositional symbol in each world. The satisfaction of a wff φ in a world w of the model $\mathfrak{M}$, denoted by $\mathfrak{M}, w \models \varphi$, is defined by the following clauses:

1. $\mathfrak{M}, w \models p$ if $V(w, p) = 1$ for each $p \in PV$;
2. $\mathfrak{M}, w \models \neg\varphi$ iff $\mathfrak{M}, w \not\models \varphi$;
3. $\mathfrak{M}, w \models \varphi \wedge \psi$ iff $\mathfrak{M}, w \models \varphi$ and $\mathfrak{M}, w \models \psi$;
4. $\mathfrak{M}, w \models \varphi \vee \psi$ iff $\mathfrak{M}, w \models \varphi$ or $\mathfrak{M}, w \models \psi$;
5. $\mathfrak{M}, w \models \varphi \supset \psi$ iff $\mathfrak{M}, w \not\models \varphi$ or $\mathfrak{M}, w \models \psi$;
6. $\mathfrak{M}, w \models \Diamond\varphi$ iff there exists $(w, u) \in \alpha$ such that $\mathfrak{M}, u \models \varphi$.

Propositional multi-modal logic (PMML) is an extension of PML that allows more than one modality [16]. PMML is particularly suitable for reasoning about relational structures that contain a number of binary relations, e. g., social networks. The alphabet of PMML consists of a set of propositional symbols PV, a set of relational symbols REL, the Boolean connectives, the relational converse symbol $^-$, and the modality-forming symbol $\langle\rangle$. The set of wffs of PMML is the smallest set containing PV that satisfies the following conditions:

- if φ is a wff, then $\neg\varphi$ is a wff;
- if φ is a wff and r is a relational symbol, then $\langle r\rangle\varphi$ and $\langle r^-\rangle\varphi$ are wffs;
- if φ and ψ are wffs, then $\varphi \wedge \psi$, $\varphi \vee \psi$, and $\varphi \supset \psi$ are wffs.

We abbreviate $\neg\langle r\rangle\neg\varphi$ and $\neg\langle r^-\rangle\neg\varphi$ as $[r]\varphi$ and $[r^-]\varphi$, respectively. A Kripke model for PMML is $\mathfrak{M} = (W, (\alpha_r)_{r \in REL}, V)$, where W and V are the same as above, and for each $r \in REL$, α_r is a binary relation on W. The satisfaction of PMML wffs is defined in the same way as that of PML wffs, except that clause 6 is replaced by

6. $\mathfrak{M}, w \models \langle r\rangle\varphi$ iff there exists $(w, u) \in \alpha_r$ such that $\mathfrak{M}, u \models \varphi$;
7. $\mathfrak{M}, w \models \langle r^-\rangle\varphi$ iff there exists $(u, w) \in \alpha_r$ such that $\mathfrak{M}, u \models \varphi$.

Note that we need the converse modalities $\langle r^-\rangle$ because all positional equivalences are defined with respect to their in-neighborhoods and out-neighborhoods. The modalities $\langle r\rangle$ allow us to access the out-neighborhood of a world, whereas $\langle r^-\rangle$ is needed to access the in-neighborhood.

Graded Modal Logic

Graded modal logic (GML) is a generalization of PMML that allows us to consider the number of worlds accessible from a given world [33]. The alphabet of GML is the same as that of PMML; however, the second formation rule of wffs is replaced by the following rule:

- if φ is a wff and r is a relational symbol, then $\langle r\rangle_n\varphi$ and $\langle r^-\rangle_n\varphi$ are wffs for each natural number n.

We also write $[r]_n\varphi$ and $[r^-]_{\varphi}$ as the abbreviations of $\neg\langle r\rangle_n\neg\varphi$ and $\neg\langle r^-\rangle_n\neg\varphi$, respectively.

The Kripke models for GML are the same as those for PMML, but clauses 6 and 7 for the conditions of satisfaction of PMML wffs are modified as follows:

6. $\mathfrak{M}, w \models \langle r\rangle_n\varphi$ iff $|\{u \in W \mid (w, u) \in \alpha_r$ and $\mathfrak{M}, u \models \varphi\}| > n$;
7. $\mathfrak{M}, w \models \langle r^-\rangle_n\varphi$ iff $|\{u \in W \mid (u, w) \in \alpha_r$ and $\mathfrak{M}, u \models \varphi\}| > n$.

According to the semantics, the modalities $\langle r\rangle_0$ and $\langle r^-\rangle_0$ in GML are respectively equivalent to $\langle r\rangle$ and $\langle r^-\rangle$ in PMML. Intuitively, $\mathfrak{M}, w \models \langle r\rangle_n\varphi$ means that φ is true in more than n worlds that are r-accessible from w, whereas $\mathfrak{M}, w \models [r]_n\varphi$ means that φ is false in no more than n worlds that are r-accessible from w. Thus, we abbreviate $\langle r\rangle_{n-1}\varphi \wedge [r]_n\neg\varphi$ (resp. $\langle r^-\rangle_{n-1}\varphi \wedge [r^-]_n\neg\varphi$) as $(r)_n\varphi$ (resp. $(r^-)_n\varphi$) to denote that there are exactly n r-accessible (resp. r^--accessible) worlds that satisfy φ.

Hybrid Logic

Hybrid logics (HL) are extensions of modal logics that include symbols to name possible worlds in models. The development of HL dates back to A. Prior's 1951 work, which had a philosophical foundation [29,30] (see [1] for a brief overview). The main idea is to use a special kind of atomic formula, called a nominal, to refer to possible worlds in models. The idea was adopted by the Sofia school in the late 80s and early 90s [14,24,25]. Subsequently, Blackburn and Seligman, who have been influential in the recent development of HL, adopted it in the late 90s [4,6].

Here, we only need the simplest hybrid logic (SHL). The alphabet of SHL is the same as that of PMML plus a set of nominals NOM that is disjoint with the set PV. The set of wffs of SHL is the smallest set containing $PV \cup NOM$ that satisfies the formation rules of PMML. The key semantic idea of HL is that each nominal must be true in exactly one possible world in any model. In other words, a nominal names a world by being true in that world and nowhere else. The idea can be easily formalized by extending PMML models with a component that assigns a nominal to a possible world. A hybrid model is then $\mathfrak{M} = (W, (\alpha_r)_{r \in REL}, V, f)$, where $(W, (\alpha_r)_{r \in REL}, V)$ is a PMML model and $f\colon NOM \to W$ is the mapping that assigns a nominal to a possible world. The satisfaction of SHL wffs is defined by clauses 1–7 for PMML, with the addition of the following clause for nominals:

8. $\mathfrak{M}, w \models i$ if $f(i) = w$ for each $i \in NOM$.

From Relational Granulation to Functional Granulation

In this section, we explain how to transform positional equivalences from relational granulation to functional granulation. Given a social network $\mathfrak{N} = (A, (P_i)_{i \in I}, (\alpha_j)_{j \in J})$, the steps are as follows:

1. For each kind of equivalence, find a corresponding modal language based on $\mathfrak{N}$. The three kinds of positional equivalence discussed earlier correspond to the following logical languages:
   (a) Structural equivalence: SHL
   (b) Regular equivalence: PMML
   (c) Exact equivalence: GML.
2. Consider $\mathfrak{N}$ as a model in the corresponding logic by a simple transformation. The set of actors is the set of possible worlds in the resultant model.
3. Find a sublanguage $\mathcal{L}$ of the corresponding logic. Prove that for any two actors x and y, $x \cong y$ iff x and y satisfy the same set of wffs in $\mathcal{L}$, where $\cong$ is $\cong_s$, $\cong_r$, or $\cong_e$ (see Fig. 5).

Structural Equivalences and Hybrid Logics

We use SHL to characterize structural equivalences. Given a social network $\mathfrak{N} = (A, (P_i)_{i \in I}, (\alpha_j)_{j \in J})$, we define an SHL language with the following alphabet:

1. $PV = \{p_i \mid i \in I\}$;
2. $REL = J$;
3. $NOM = A$.

The social network $\mathfrak{N}$ is transformed into a hybrid model $\mathfrak{M}_{\mathfrak{N}} = (A, (\alpha_j)_{j \in J}, V, id)$, where V is defined by $V(x, p_i) = 1$ iff $x \in P_i$ for all $x \in A$ and $i \in I$, and id is the identity function on A. Let us now define a sublanguage $\mathcal{H}_{\mathfrak{N}}$ of the SHL as the set of wffs $PV \cup \{\langle j \rangle a, \langle j^- \rangle a \mid j \in REL, a \in NOM\}$. We say that two actors, x and y, are $\mathcal{H}_{\mathfrak{N}}$-equivalent with respect to $\mathfrak{N}$ if for all $\varphi \in \mathcal{H}_{\mathfrak{N}}$, ($\mathfrak{M}_{\mathfrak{N}}, x \models \varphi$ iff $\mathfrak{M}_{\mathfrak{N}}, y \models \varphi$). Then, we have the following characterization theorem.

Theorem 1 *Let $\mathfrak{N} = (A, (P_i)_{i \in I}, (\alpha_j)_{j \in J})$ be a social network; then, for all $x, y \in A$, $x \cong_s y$ in $\mathfrak{N}$ iff x and y are $\mathcal{H}_{\mathfrak{N}}$-equivalent with respect to $\mathfrak{N}$.*

Proof By the satisfaction condition, ($\mathfrak{M}_{\mathfrak{N}}, x \models \langle j \rangle a$ iff $(x, a) \in \alpha_j$. Thus, x and y are $\mathcal{H}_{\mathfrak{N}}$-equivalent iff the following two conditions are satisfied

1. $x \in P_i$ iff $y \in P_i$ for all $i \in I$ (as $PV \subseteq \mathcal{H}_{\mathfrak{N}}$),
2. for all $j \in J$ and $a \in A$, $(x, a) \in \alpha_j$ iff $(y, a) \in \alpha_j$ and $(a, x) \in \alpha_j$ iff $(a, y) \in \alpha_j$ (since both $\langle j \rangle a, \langle j^- \rangle a$ and $\langle j \rangle a, \langle j^- \rangle a$ are in $\mathcal{H}_{\mathfrak{N}}$).

These two conditions are exactly the same as those in Definition 1. Since $x \cong_s y$ iff (x, y) is in the coarsest structural equivalence, $x \cong_s y$ iff these two conditions are satisfied.□

Regular Equivalences and Multimodal Logics

The first logical characterization of positional equivalences is that of regular equivalences by dynamic logic – a particular type of PMML. This is shown in [22] by the connection between regular equivalence and the well-known notion of bisimulation in modal logics. We present the result here by following the general procedure of the transformation from relational granulation to functional granulation.

Given a social network $\mathfrak{N} = (A, (P_i)_{i \in I}, (\alpha_j)_{j \in J})$, we define a PMML language with the following alphabet:

1. $PV = \{p_i \mid i \in I\}$;
2. $REL = J$.

The social network $\mathfrak{N}$ is transformed into a Kripke model $\mathfrak{M}_{\mathfrak{N}} = (A, (\alpha_j)_{j\in J}, V)$, where V is defined by $V(x, p_i) = 1$ iff $x \in P_i$ for all $x \in A$ and $i \in I$. Let $\mathcal{L}_{\mathfrak{N}}$ be the set of wffs of the PMML and $\mathcal{L}_{\mathfrak{N}}$-equivalence be defined the same as $\mathcal{H}_{\mathfrak{N}}$-equivalence.

Theorem 2 ([22]) *Let $\mathfrak{N} = (A, (P_i)_{i\in I}, (\alpha_j)_{j\in J})$ be a social network; then, for all $x, y \in A$, $x \cong_r y$ in $\mathfrak{N}$ iff x and y are $\mathcal{L}_{\mathfrak{N}}$-equivalent with respect to $\mathfrak{N}$.*

Proof It is shown in [22] that regular equivalences are exactly bisimulations, i. e., $x \cong_r y$ iff x and y are bisimular. By Theorems 2.20 and 2.24 in [5], x and y are bisimular iff they are $\mathcal{L}_{\mathfrak{N}}$-equivalent with respect to $\mathfrak{N}$. □

Exact Equivalences and Graded Modal Logics

Following the procedure used in the previous subsection, we define $\mathcal{G}_{\mathfrak{N}}$ as the set of wffs of the GML based on the given network $\mathfrak{N}$. Then, we have the following theorem.

Theorem 3 *Let $\mathfrak{N} = (A, (P_i)_{i\in I}, (\alpha_j)_{j\in J})$ be a social network; then, for all $x, y \in A$, $x \cong_e y$ in $\mathfrak{N}$ iff x and y are $\mathcal{G}_{\mathfrak{N}}$-equivalent with respect to $\mathfrak{N}$.*

Proof

1. ($\Rightarrow$): Assume that $x \cong_e y$; then, there exists an exact equivalence ρ on $\mathfrak{N}$ such that $(x, y) \in \rho$. We show that $\mathfrak{M}_{\mathfrak{N}}, x \models \varphi$ implies $\mathfrak{M}_{\mathfrak{N}}, y \models \varphi$ by induction on the complexity of φ (the proof of the converse direction is exactly the same). The case where $\varphi \in PV$ follows from condition 1 of Definition 3. The cases for Boolean connectives are straightforward by the induction hypothesis. For formulas of the form $\langle j\rangle_n\psi$ or $\langle j^-\rangle_n\psi$, we prove the case for $\langle j\rangle_n\psi$ by using the condition $[\![N^+_{\alpha_j}(x)]\!]_\rho = [\![N^+_{\alpha_j}(y)]\!]_\rho$. The case for $\langle j^-\rangle_n\psi$ is proved similarly by applying $[\![N^-_{\alpha_j}(x)]\!]_\rho = [\![N^-_{\alpha_j}(y)]\!]_\rho$. Now, if $\mathfrak{M}_{\mathfrak{N}}, x \models \langle j\rangle_n\psi$, then there exist more than n worlds in $N^+_{\alpha_j}(x)$ that satisfy ψ. Assume $u_0, u_1, \ldots, u_n$ are among these worlds. Since $[\![N^+_{\alpha_j}(x)]\!]_\rho = [\![N^+_{\alpha_j}(y)]\!]_\rho$, there exist $u'_0, u'_1, \ldots, u'_n \in N^+_{\alpha_j}(y)$ such that $(u_i, u'_i) \in \rho$ for $0 \le i \le n$. By the induction hypothesis, $\mathfrak{M}_{\mathfrak{N}}, u'_i \models \psi$ for $0 \le i \le n$; thus, $\mathfrak{M}_{\mathfrak{N}}, y \models \langle j\rangle_n\psi$.
2. ($\Leftarrow$): Let us define an equivalence relation ρ on A by

$$(x, y) \in \rho$$
$$\text{iff} \quad x \text{ and } y \text{ are } \mathcal{G}_{\mathfrak{N}}\text{-equivalent with respect to } \mathfrak{N}.$$

We prove that ρ is an exact equivalence. The first condition of Definition 3 is straightforward. For the second condition, let us assume that $(x, y) \in \rho$, $[\![N^+_{\alpha_j}(x)]\!]_\rho = ([A]_\rho, m_1)$, and $[\![N^+_{\alpha_j}(y)]\!]_\rho = ([A]_\rho, m_2)$ for some $j \in J$. Assume also that $[A]_\rho = \{[z_i]_\rho \mid 1 \le i \le k\}$. First, we consider z_1 as an example. For each $2 \le i \le k$, since $[z_1]_\rho$ is an equivalence class distinct from $[z_i]_\rho$, there exists a $\mathcal{G}_{\mathfrak{N}}$-wff ψ_i such that $\mathfrak{M}_{\mathfrak{N}}, z_i \models \psi_i$, but $\mathfrak{M}_{\mathfrak{N}}, z_1 \models \neg\psi_i$. If $m_1([z_1]_\rho) = n_1 > 0$, then there exist exactly n_1 α_j-successors of x that are $\mathcal{G}_{\mathfrak{N}}$-equivalent to z_1. Thus, $\mathfrak{M}_{\mathfrak{N}}, x \models (j)_{n_1}(\psi \wedge \bigwedge_{i=2}^k \neg\psi_i)$ for all ψ satisfied in z_1. Since $(x, y) \in \rho$, it is that $\mathfrak{M}_{\mathfrak{N}}, y \models (j)_{n_1}(\psi \wedge \bigwedge_{i=2}^k \neg\psi_i)$ for all ψ satisfied in z_1. This means there are exactly $n_1\alpha_j$-successors of y that satisfy $\bigwedge_{i=2}^k \neg\psi_i$ (by letting $\psi = \top$). Thus, these n_1 α_j-successors of y must satisfy all ψ satisfied in z_1 and be $\mathcal{G}_{\mathfrak{N}}$-equivalent to z_1. Since no other α_j-successors of y can be $\mathcal{G}_{\mathfrak{N}}$-equivalent to z_1 (otherwise there would be more than $n_1\alpha_j$-successors of y satisfying $\bigwedge_{i=2}^k \neg\psi_i$), $m_2([z_1]_\rho) = n_1$. In the same way, we can prove that if $m_2([z_1]_\rho) > 0$, then $m_1([z_1]_\rho) > 0$. Thus, $m_1([z_1]_\rho) = m_2([z_1]_\rho)$. Similarly, we can prove $m_1([z_i]_\rho) = m_2([z_i]_\rho)$ for $2 \le i \le k$. Consequently, we have $m_1 = m_2$ and $[\![N^+_{\alpha_j}(x)]\!]_\rho = [\![N^+_{\alpha_j}(y)]\!]_\rho$. The proof of $[\![N^-_{\alpha_j}(x)]\!]_\rho = [|N^-_{\alpha_j}(y)|]_\rho$ is analogous, except that the modality $(j)_{n_1}$ is replaced by $(j^-)_{n_1}$. □

Observational Equivalences

We have presented the logical characterizations of three kinds of positional equivalence. The theoretical implication of these characterizations is that positional equivalences based on relational granulation can be transformed into indiscernibility relations based on a set of binary attributes. Let $\mathfrak{N} = (A, (P_i)_{i\in I}, (\alpha_j)_{j\in J})$ be a social network and let F denote $\mathcal{H}_{\mathfrak{N}}$, $\mathcal{L}_{\mathfrak{N}}$, or $\mathcal{G}_{\mathfrak{N}}$. Then, (A, F) is a (generalized) data table such that IND(F) is the corresponding positional equivalence. However, from a practical viewpoint, since both $\mathcal{L}_{\mathfrak{N}}$ and $\mathcal{G}_{\mathfrak{N}}$ are infinite sets, it is generally difficult (if not impossible) to decide IND(F) based on such a data table representation. This motivates us to define a new kind of positional equivalence based on observations.

Let $\mathfrak{N}$ be a social network and F be a *finite* set of wffs in an appropriate logic language. Then, two actors, x and y, are *F-observationally equivalent*, denoted by $x \cong_F y$, if for any wff $\varphi \in F$, x satisfies φ iff y satisfies φ. The wffs in F denote the observable properties of the actors in the network. Due to the limited capability of the observer, the set of observable properties is usually finite. Depending on the observable properties chosen, we can focus on different parts of the social network. For example, positional analysis for a genealogical study and a political investigation may have to consider different relationships between actors. As shown by the above results, the choice of an appropriate logic language is also crucial, since it may result in different notions of positional equivalence.

A recent approach to structural equivalence that conforms with our definition of observational equivalence is reported in [28]. In that approach, a set of institutions E is given and each actor may belong to several institutions. Two actors are structurally equivalent if they belong to the same set of institutions. To avoid confusion with the structural equivalence defined in Sect. "Positional Analysis – Relational Granulation", we call the equivalence relation defined in [28] *institutional equivalence*. Note that institutional equivalence is in fact the same as structural equivalence defined in Sect. "Positional Analysis – Relational Granulation", if we consider the universe as the union of A and E and represent the network as a bipartite graph. A place (social position) is then defined as an equivalence class of actors (or equivalently, the set of institutions that those actors belong to). Here, our set of observation sentences corresponds to the set of institutions in [28]. While institutions are primitive entities, our observation sentences are formed from the social relations and attributes of actors. In fact, a wff in F can also be regarded as an intensional definition of an institution. In this sense, observational equivalence is a kind of intensionally institutional equivalence.

Conclusion and Future Directions

This article starts with an introduction to rough set theory – the basic theory of functional granulation. We then review several notions of positional equivalence in social network analysis and consider them as relational granulation from the GrC perspective. By using modal logics as a bridge, we transform the relational granulation-based positional equivalences into functional granulation. However, the transformation is mainly theoretically interesting, since the set of functional attributes used to granulate a social network may be infinite. This practical consideration motivates us to propose the notion of observational equivalence. It conforms with an alternative definition of structural equivalence, which we call institutional equivalence to avoid confusion.

More generally, both observational equivalence and institutional equivalence can be analyzed in the framework of formal concept analysis (FCA) [13]. FCA is a conceptual knowledge representation and data analysis method in which a *formal context* is defined as a triple (A, F, R), where A and F are sets of objects and attributes respectively; and $R \subseteq A \times F$ is a binary relation between A and F, called the incidence of the formal context. For $A' \subseteq A$ and $F' \subseteq F$, let us denote $R(A') = \{f \in F \mid \forall a \in A', (a, f) \in R\}$ and $R^{-}(F') = \{a \in A \mid \forall f \in F', (a, f) \in R\}$. Then, a *formal concept* is defined as a pair (A', F') such that $A' \subseteq A$, $F' \subseteq F$, $R(A') = F'$ and $R^{-}(F') = A'$. Here, we call A' the extent and F' the intent of the formal concept. The application of FCA to sociology has been proposed in [12]. In that application, the formal context (also called the Galois lattice) is used to model two-mode network data, where the set of actors is a mode of the data, the set of events is the other mode of data, and the relation between the two modes is the participation relationship between the actors and events. When we regard A as the set of actors, F as the set of observation sentences (resp. institution), and R as the satisfaction (resp. membership) relation, a formal context can be constructed out of a social network. However, a social position (or place) defined by observational or institutional equivalences may not be the extent of a formal concept. Thus, it is tempting to define a social position alternatively as a formal concept of such formal contexts. Unlike other positional analysis techniques that define a social position as an equivalence class of some positional equivalence, the extents of different formal concepts may overlap. This means that social positions may not be disjoint in such a definition. The theoretical and practical implications of such positional analysis requires further research.

Another future research direction concerns the practical computation of positional equivalences. As mentioned earlier, when a social network is transformed into a data table (A, F) for positional analysis and F is infinite, it is difficult to decide if two actors are $\mathrm{IND}(F)$-equivalent by exhaustively comparing the satisfaction of all wffs in x and y. While observational equivalence provides a practical approximation of the computation, we would like to know if precise computation is possible. To answer this question, the inference of the *most specific concept* (msc) in description logics [2] may be helpful. Description logics are a family of knowledge representation formalisms that have strong connections with modal logics [3,23]. Indeed, the description languages $\mathcal{ALC}$ and $\mathcal{ALCN}$ correspond to the sublanguages of PMML and GML respectively [3,23]. Given a network represented in a description language, the msc of an actor x, denoted by $msc(x)$, corresponds to a wff that satisfies x and implies any wffs that satisfy x. Thus, if $msc(x)$ and $msc(y)$ are logically equivalent, then x and y will satisfy the same set of wffs, i. e., $(x, y) \in \mathrm{IND}(F)$. The problem for researchers is to find logical languages that can characterize the aforementioned positional equivalences adequately and whose msc problems can be solved effectively.

In summary, we have proposed a logical approach to positional analysis in social networks from a GrC perspective. The approach facilitates a connection between more complicated relational granulation and simpler functional granulation. While there remain problems to be solved in the research programme, the approach seems promising for cross-fertilization between the fields of GrC and SNA.

Bibliography

1. Areces C, ten Cate B (2007) Hybrid logics. In: Blackburn P, Van Benthem J, Wolter F (eds) Handbook of Modal Logic. Elsevier, Amsterdam, pp 821–868
2. Baader F, Küsters R (2007) Nonstandard inferences in description logics: the story so far. In: Gabbay DM, Goncharov SS, Zakharyaschev M (eds) Mathematical Problems from Applied Logic I: Logics for XXIst Century. Springer, Berlin
3. Baader F, Nutt W (2002) Basic description logics. In: Baader F, Calvanese D, McGuinness DL, Nardi D, Patel-Schneider PF (eds) Description Logic Handbook. Cambridge University Press, Cambridge, pp 47–100
4. Blackburn P (2000) Representation, reasoning, and relational structures: a hybrid logic manifesto. Logic J IGPL 8(3):339–625
5. Blackburn P, de Rijke M, Venema Y (2001) Modal Logic. Cambridge University Press, Cambridge
6. Blackburn P, Seligman J (1995) Hybrid languages. J Log Lang Inf 4:251–272
7. Borgatti SP, Everett MG (1989) The class of all regular equivalences: algebraic structure and computation. Soc Netw 11(1):65–88
8. Boyd JP, Everett MG (1999) Relations, residuals, regular interiors, and relative regular equivalence. Soc Netw 21(2):147–165
9. Chellas BF (1980) Modal Logic: An Introduction. Cambridge University Press, Cambridge
10. Everett MG, Borgatti SP (1994) Regular equivalences: general theory. J Math Sociol 18(1):29–52
11. Fan TF, Liau CJ, Liu DR, Tzeng GH (2006) Granulation based on hybrid information systems. In: Proc of the 2006 IEEE International Conference on Systems Man and Cybernetics, Taipei, pp 4768–4772
12. Freeman LC, White DR (1993) Using Galois lattices to represent network data. Sociol Methodol 23:127–146
13. Ganter B, Wille R (1999) Formal Concept Analysis: Mathematical Foundations. Springer, Berlin
14. Gargov G, Goranko V (1993) Modal logic with names. J Philos Log 22(6):607–636
15. Hanneman RA, Riddle M (2005) Introduction to Social Network Methods. University of California, Riverside
16. Horrocks I, Hustadt U, Sattler U, Schmidt R (2007) Computational modal logic. In: Blackburn P, Van Benthem J, Wolter F (eds) Handbook of Modal Logic. Elsevier, Amsterdam, pp 181–245
17. Kripke S (1959) A completeness theorem in modal logic. J Symb Log 24(1):1–14
18. Kripke S (1963) Semantic analysis of modal logic I: normal propositional calculi. Z Math Log Grundl Math 9:67–96
19. Lerner J (2005) Role assignments. In: Brandes U, Erlebach T (eds) Network Analysis. LNCS, vol 3418. Springer, Berlin, pp 216–252
20. Liau CJ, Lin TY (2005) Reasoning about relational granulation in modal logics. In: Proc of the First IEEE International Conference on Granular Computing, Beijing, pp 534–558
21. Lorrain F, White HC (1971) Structural equivalence of individuals in social networks. J Math Sociol 1:49–80
22. Marx M, Masuch M (2003) Regular equivalence and dynamic logic. Soc Netw 25(1):51–65
23. Nardi D, Brachman RJ (2002) An introduction to description logics. In: Baader F, Calvanese D, McGuinness DL, Nardi D, Patel-Schneider PF (eds) Description Logic Handbook. Cambridge University Press, Cambridge, pp 5–44
24. Passy S, Tinchev T (1985) Quantifiers in combinatory PDL: completeness, definability, incompleteness. In: Fundamentals of Computation Theory FCT 85. LNCS, vol 199. Springer, Berlin, pp 512–519
25. Passy S, Tinchev T (1991) An essay in combinatory dynamic logic. Inf Comput 93:263–332
26. Pawlak Z (1982) Rough sets. Int J Inf Comput Sci 11(15):341–356
27. Pawlak Z (1991) Rough Sets – Theoretical Aspects of Reasoning about Data. Kluwer, Dordrecht
28. Pizarro N (2007) Structural identity and equivalence of individuals in social networks: Beyond duality. Int Sociol 22(6):767–792
29. Prior A (1967) Past, Present and Future. Oxford University Press, London
30. Prior A (1968) Now. Nous 2:101–119
31. Scott J (2000) Social Network Analysis: A Handbook, 2nd edn. SAGE Publications, London
32. Sierpinski W, Krieger C (1956) General Topology. University of Toronto Press, Toronto
33. van der Hoek W (1992) On the semantics of graded modalities. J Appl Non-Class Log 2(1):81–123
34. Wasserman S, Faust K (1994) Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge
35. White DR, Reitz KP (1983) Graph and semigroup homomorphisms on netwoks and relations. Soc Netw 5(1):143–234
36. Zadeh LA (1979) Fuzzy sets and information granularity. In: Gupta N, Ragade R, Yager R (eds) Advances in Fuzzy Set Theory and Applications. North-Holland, Amsterdam, pp 3–18
37. Zadeh LA (1997) Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets Syst 19:111–127

Social Network Visualization, Methods of

LINTON C. FREEMAN
Department of Sociology and Institute for Mathematical Behavioral Science, School of Social Sciences,
University of California, Irvine, USA

Article Outline

Glossary

Adjacent A node is adjacent to another if there is an edge connecting them.

Arrow A line directed from one node to another.

Binary relation A two valued yes/no or on/off relation.

Bipartite graph A graph, $B = \langle N, E \rangle$ where N is a finite set of nodes and E is a collection of pairs of nodes in which N is partitioned into two disjoint subsets, N_1 and N_2, and no edge in E has both end points in the same subset.

Blockmodeling A procedure for clustering actors such that the actors in each cluster share similar patterns of ties both within and between clusters.

Connected Any two nodes in a graph are said to be connected if there is a path from one to the other; a graph is connected if there is a path connecting every pair of nodes.

Cycle Any path that begins and ends at the same node.

Digraph A directed graph.

Directed graph A graph $D = \langle N, A \rangle$ where N is a finite collection of nodes and A is a set of pairs linked by directed lines or arrows.

Directed line A line going from a node to another representing a non-reciprocated link.

Edge A line connecting two nodes representing a reciprocated link.

Edge labeled graph A graph in which at least two kinds of connections between nodes are identified.

Formal concept analysis A method of data analysis based on Galois lattice structure.

Galois lattice A dual structure that displays the dependencies of both objects and their properties.

Geodesic The shortest path between two nodes.

Graph A graph $G = \langle N, E \rangle$ where N is a finite set of nodes and E is a collection of pairs of nodes represented as edges.

Hyperedge An edge in a hypergraph that can enclose more than two nodes.

Hypergraph A hypergraph, $F = \langle N, H \rangle$, consists of a set of nodes N and a collection of hyperedges, H.

Indegree The indegree of a node is the number of directed lines it receives.

Irreflexive A relation in which no edge connects any node with itself.

Multidimensional scaling A search procedure designed to represent an observed set of proximities or distances in a small number of dimensions.

Node A point in a graph.

One mode matrix A data matrix in which the rows and columns both represent the same objects.

Outdegree The outdegree of a node is the number of directed lines it sends out.

Path A path is a sequence of nodes and edges beginning with a node that has an edge connecting it to the next node in the sequence and so on.

Path length The length of a path connecting two nodes is the number of edges it contains.

Permutation A reordering of the rows, columns, or rows and columns of a matrix.

Principle diagonal The set of cells in a square matrix that runs from the upper left to the lower right.

Relation A collection of ordered or unordered pairs of nodes.

Singular value decomposition an algebraic procedure that decomposes a data matrix into its "basic structure".

Sociometry An early version of social network analysis introduced by Jacob Moreno and Helen Jennings.

Spring embedder A kind of multidimensional scaling based on a model in which it is assumed that nodes are connected by springs that pull and push on them.

Symmetric A relation in which if a node a is adjacent to another, b, then b is adjacent to a.

Tree A graph is a tree if it is connected and contains no cycles.

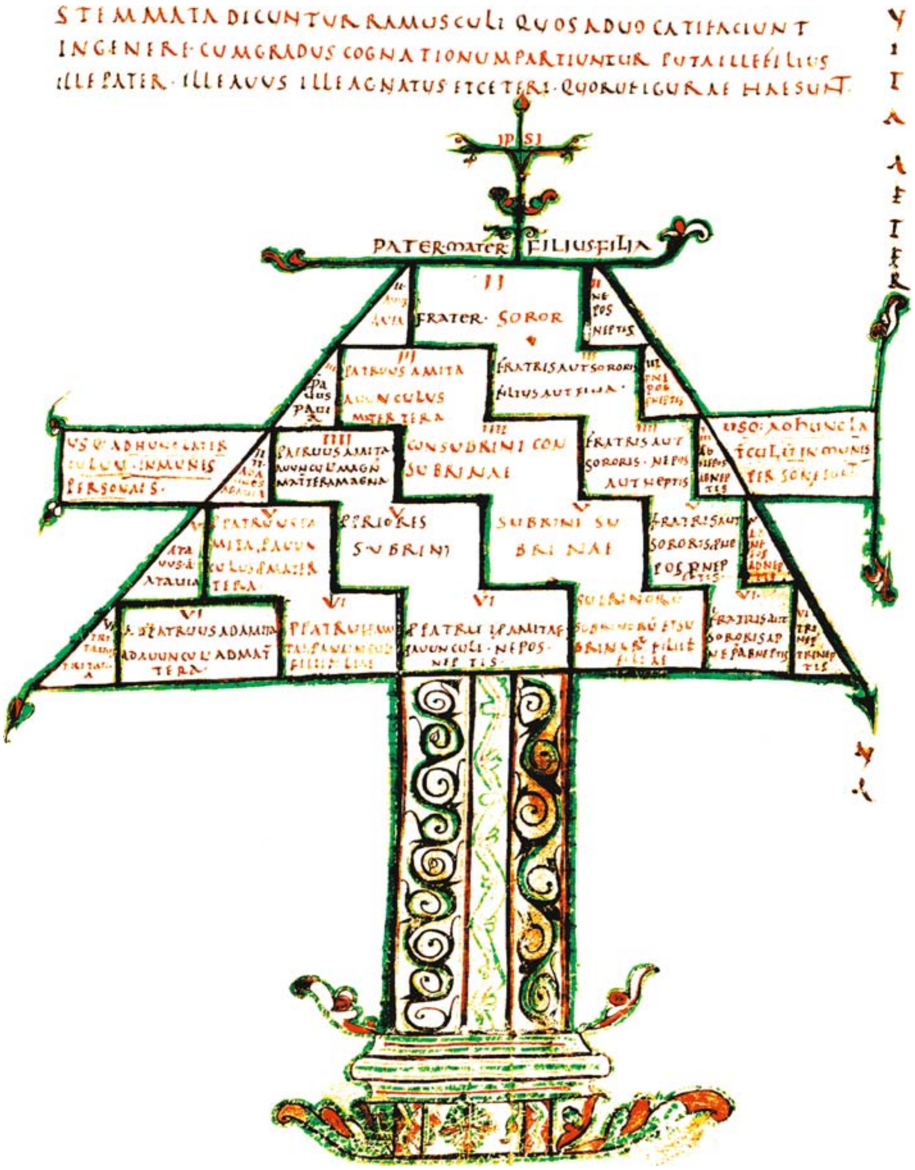
Two mode matrix A data matrix in which the rows and columns represent different objects.

Definition of the Subject

Social network visualization refers to the practice of constructing pictorial images of the connections linking social actors. The use of such images provides two benefits. It allows investigators to gain new insights into the patterning of social connections, and it helps investigators to communicate their results to others.

Introduction

Social network analysis did not emerge as a systematic field of research until early in the twentieth century [1]. But visual images of social networks were produced more than a millennium earlier. The earliest of these images that I have uncovered was produced in Spain in the middle of



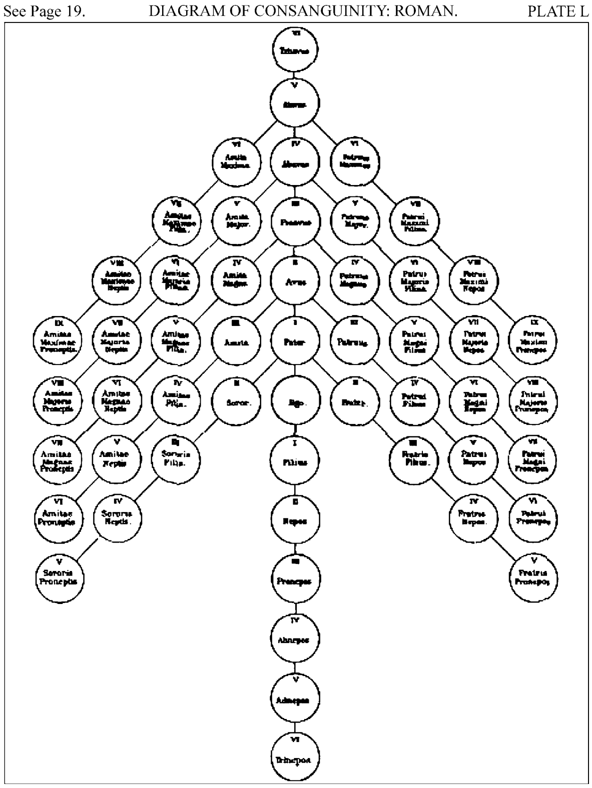
Social Network Visualization, Methods of, Figure 1
Tree of consanguinity with six degrees of relationship

the ninth century. That image is attributed to the prolific writer and Roman Catholic Saint, Isidore de Séville. It is reproduced here as Fig. 1.

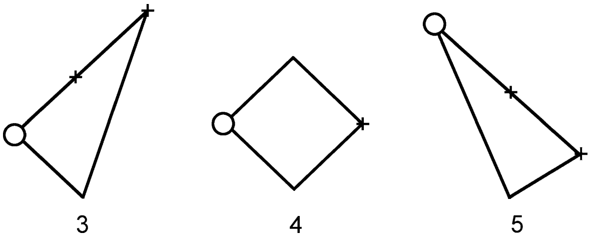
The image shown in Fig. 1 displays relationships based on genealogical descent. From the earliest times, people have been interested in kinship ties – in who is related to whom. This interest is evident in the descent lists found in the Christian bible and in the oral genealogies that were required to be memorized by Hawaiian nobles [2].

The fact that Fig. 1 takes the form of a tree shows that as early as the ninth century people saw the analogy between the branching structure of descent and that of trees. This notion was captured in a mathematical formalization in 1857 by Arthur Cayley [3]. Cayley defined a tree in mathematical graph theoretic terms. Biggs, Lloyd and Wilson (p. 38 in [4]) characterized Cayley's definition by saying that his "… use of the word 'tree' in this context is presumably derived from the diagrammatic form of these graphs, and is akin to the traditional use of the word in describing genealogical or 'family' trees."

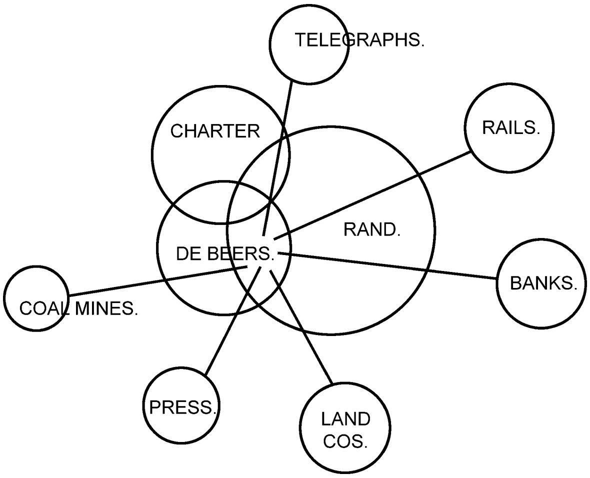
The use of trees to depict descent was, of course, continued. As time passed, however, their form became sim-



Social Network Visualization, Methods of, Figure 2
Descent in ancient Rome



Social Network Visualization, Methods of, Figure 3
Macfarlane's images of two-step marriage prohibitions



Social Network Visualization, Methods of, Figure 4
Hobson's image of corporate interlocks

plified. Lewis Henry Morgan [5] was an attorney and an anthropologist. He was interested in comparing how different peoples reckoned kinship and in 1871 he published a mammoth work containing a collection of kinship trees. Each tree depicted descent as conceived by a society somewhere in the world. Morgan's trees are quite simple. Figure 2 shows descent as it was reckoned in ancient Rome.

Twelve years later a mathematician-physicist, Alexander Macfarlane [6], produced a different kind of graphic image based still on kinship. Macfarlane set out to examine British marriage prohibitions and he represented them both algebraically and visually. His visual images depict males using plus signs (+) and females with circles (o). Earlier generations he placed higher on the page. Descent is shown by lines connecting points. A short line crossing a descent line indicates another person, of either sex, in an intermediate generation. And the lowest point is always the prohibited offspring.

The illustration shown in Fig. 3 displays all the two-step marriage relations that are prohibited by British law. The left image shows that a male may not marry his granddaughter. The middle image shows that he may not marry his sister. And the right image shows that he may not marry his grandmother. Or, put the other way, a woman may not marry her grandfather, her brother or her grandson.

Macfarlane's paper also included algebraic expressions that captured all of the same marriage prohibitions. But Sir Francis Galton [7], who attended Macfarlane's presentation, declared that his "diagrammatic form" seemed "the most distinctive and self-explanatory" of the two treatments.

Finally, in 1894, John Hobson [8] produced a visual image of a social network that was not based on kinship. He had collected two mode (corporation by director) data on interlocking corporate directorates. He reasoned that, to the degree that corporations shared directors, they could be expected to cooperate and work together.

Hobson's illustration was designed to show the interlock among, as he put it, "the small inner ring of South African finance." Corporations are depicted as circles, and interlock is shown by overlapping or by a line connecting two circles. Hobson's image is reproduced here as Fig. 4.

The important feature of this image is that it displays a connection linking more than two corporations. Hobson's data showed that three corporations, Charter, Rand and De Beers, all shared directors in common. And, at the same time, Rand and De Beers also both shared directors with coalmines, telegraphs, rails, and others. The overlaps in his image allowed him to display which companies shared with which others.

It is clear, then, that a concern with connections among social actors and the use of visual images have a long history of intimate association. It should come as no surprise therefore that images played an important part in the development of social network analysis when it did emerge as an organized field of research.

Visualization in Social Network Analysis

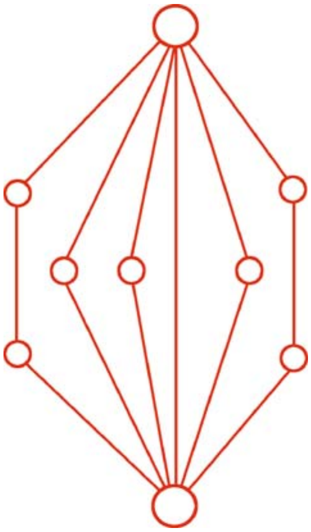
In the book cited above (p. 3 in [1]), I described the modern science of social network analysis as possessing four defining properties. They were:

1. It embodies ideas about the importance of social ties linking social actors.
2. It collects data reflecting those ties.
3. It involves the use of graphic imagery.
4. It employs mathematical and/or computational models.

Pre-network research often included one or two of those properties, but in the late 1920s each of two independent research teams came up with efforts that included all four.

One took place in the early 1930s. It involved a psychiatrist, Jacob L. Moreno, and a psychologist, Helen H. Jennings. Together, they developed an approach they called "sociometry." They reported two huge studies, both focused on examining the structure of social ties. One was conducted among prisoners at Sing Sing Prison in Ossining, New York [9] and the other among young delinquents at the New York State Training School for Girls in Hudson, New York [10].

Both Moreno–Jennings studies involved the extensive use of graphic images. The image shown in Fig. 5 was included in their report on the research at Sing Sing prison (Moreno, 1932). In that figure, individuals or other kinds of social actors are represented as points or *nodes* and links between pairs of social actors are lines or *edges* connecting pairs of nodes. In Fig. 5 Moreno was concerned with the positions of individuals and the patterning of their ties. As he put it, the individuals at the top and the bottom were "dominant" and the image showed that those dominant individuals were linked both "directly" and "indirectly".



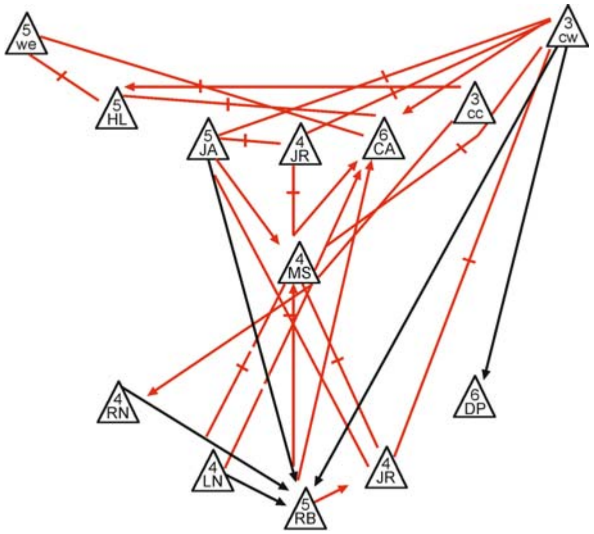
Social Network Visualization, Methods of, Figure 5
Image of a Pattern of Linkages

Most of the data collected by Moreno and Jennings involved asking individuals whom they liked or disliked. In data of that sort, choices are seldom reciprocated. So Moreno and Jennings drew lines with arrowheads to reveal who chose whom. Mutual choices were drawn without arrowheads and they also included a small line bisecting the main line connecting the two nodes.

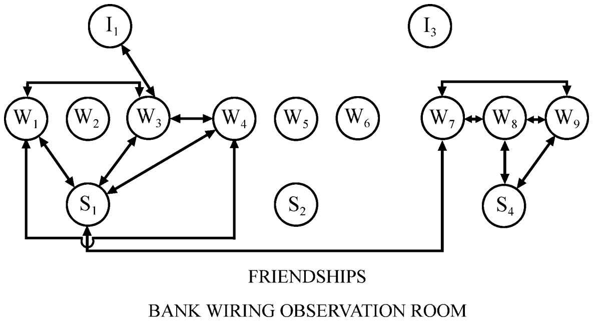
Moreno and Jennings often required subjects to report both their likes and their dislikes. By using different colors, red for likes and black for dislikes, a single image could display both. The image shown in Fig. 6 was published in the Moreno–Jennings report on the Hudson School (Moreno, 1934). It depicts positive and negative choices among 13 members of an American football team. Moreover, it contains another innovation. The various team members are placed in the drawing in approximately the same relative locations that they occupied on the football field. That arrangement shows the players' positions and permits the viewer to evaluate the impact of physical proximity on the patterning of social linkages.

Figures of the sort used by Moreno and Jennings had a major impact on the style of graphic imagery used subsequently in social network analysis. For the most part, social network analysts have represented social actors as nodes and links between actors as edges or as directed lines with arrowheads.

The second introduction of the social network approach also occurred in the early 1930s. An anthropologist, W. Lloyd Warner, and a collection of his colleagues and students at Harvard, conducted three elab-



Social Network Visualization, Methods of, Figure 6
Positive and Negative Choices in a Football Team

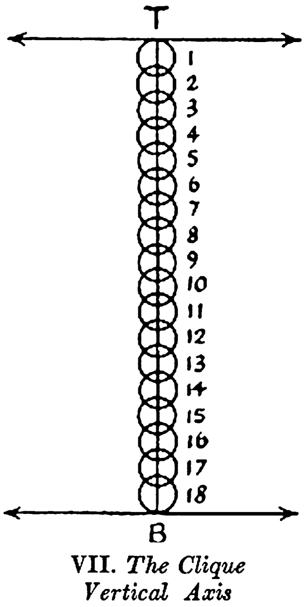


Social Network Visualization, Methods of, Figure 7
Friendships linking factory workers

orate network analytic projects. One was a study of an industrial factory, the Western Electric plant in Cicero, Illinois [11]. The other two were studies of communities: one focused on a New England town, Newburyport, Massachusetts [12], and the other on a southern town, Natchez, Mississippi [13].

The image shown in Fig. 7 was produced as part of the factory study. It displays observed friendship ties among pairs of individuals who worked together in the same workroom. It was drawn using nodes and two-headed lines instead of edges, but it is very similar to the images produced by Moreno and Jennings. In addition, as in Fig. 6 above, the impact of physical space was displayed; workers were placed in the drawing in positions that reflected the locations of their workstations.

In reporting their study of Newburyport, Warner and Lunt used the kind of drawing of overlapping circles that



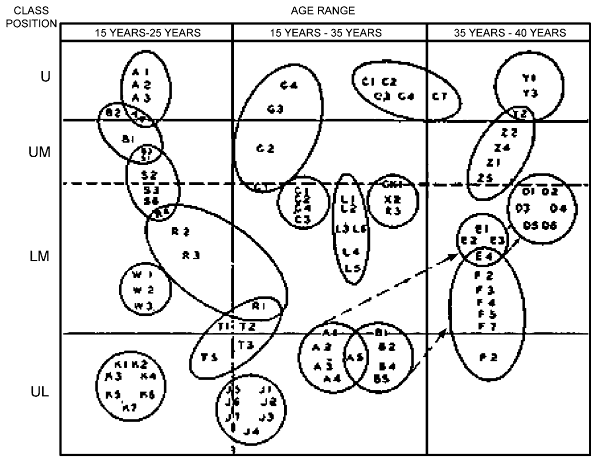
Social Network Visualization, Methods of, Figure 8
An idealized pattern of overlapping 'cliques'

Hobson had used to construct Fig. 4. But here that image was introduced, not to describe data, but to propose an idea they had about social structure. The diagram in Fig. 8 represents the investigators' idealized version of the expected structure of overlaps among subgroups in the presence of social class. The idea is that only subgroups that are close to one another in class ranking are likely to have overlapping memberships.

In their study of Natchez Davis, Gardner and Gardner [13] employed the same diagrammatic form to display two mode data reflecting on the earlier Newburyport hypothesis. Figure 9 shows subgroups of black males and their overlaps. In that image the men are arranged in terms of both social class and age. Both, it turned out, provided important bases for grouping.

Finally, in that same report, Davis, Gardner and Gardner also introduced an entirely different kind of social network image. Like Hobson, they had collected two mode data. Eighteen women were designated in the rows of their data matrix and fourteen social events were depicted in the columns. That matrix is reproduced here as Fig. 10.

The data shown in Fig. 10 were all collected during a single year. But, by examining the column headings, it is clear that Davis and his colleagues did not arrange the social events according to the dates upon which they took place. Instead, they listed both the events and the women who attended them in such a way that the arrangement itself suggests that these women were organized into two



Social stratification of a group of colored cliques: ages 15–60

Social Network Visualization, Methods of, Figure 9
Stratification, age and overlapping groups

groups. The two groups overlap, but for the most part they are distinct. Most of the women in the top half of the matrix attended the leftmost five events. And most of the women in the bottom half attended the rightmost five events. The middle four events apparently brought both groups of women together.

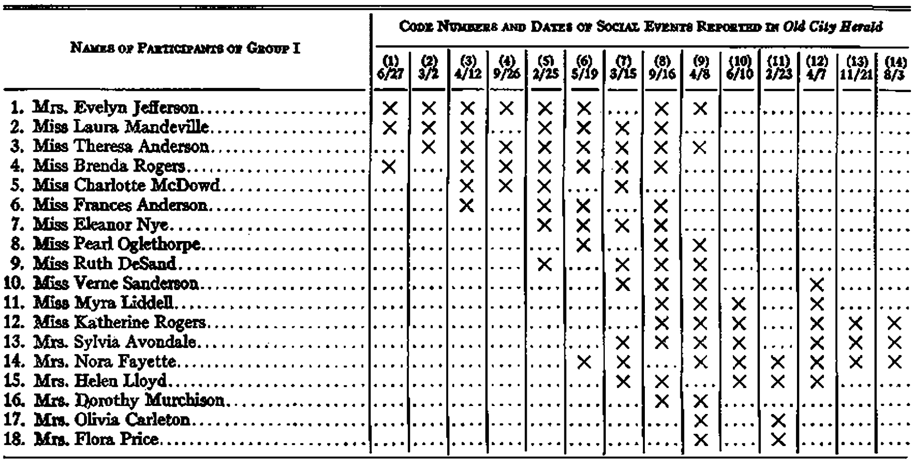
This arrangement of women and events was self-consciously produced by the authors. Davis, Gardner and Gardner were convinced that these women were organized into two groups and they presented their data matrix in a way that would illustrate that conclusion. The interesting thing is that these authors never commented explicitly about how they had rearranged the columns and rows in their matrix. They simply organized their display in a way that would make the point.

From the outset, then, four kinds of images have played important parts in the development of social network analysis. These first network graphics included drawings displaying (1) one mode undirected relations, (2) one mode directed relations, (3) two mode relations and (4) one or two mode data matrices. A few other kinds of network images have been used since then, but the four originals – particularly those based on one mode undirected and one mode directed relations – still dominate the field. In the next four sections we will examine the four original kinds of images and how their use has evolved in the social network context.

Images Based on One Mode Undirected Relations

Mathematically speaking, the node and edge images introduced by Moreno and Jennings in Fig. 5 are *graphs*. A graph $G = \langle N, E \rangle$ where N is a finite set of nodes and E is a collection of pairs of nodes. In graph visualizations, a pair of nodes in E is presented as a line connecting the two nodes in question. Two nodes are called *adjacent* if there is an edge directly connecting them to each other. A graph embodies a *binary* (yes/no or on/off) relation that is *irreflexive* (no node is adjacent to itself) and *symmetric* (if a node a is adjacent to another, b, then b is adjacent to a).

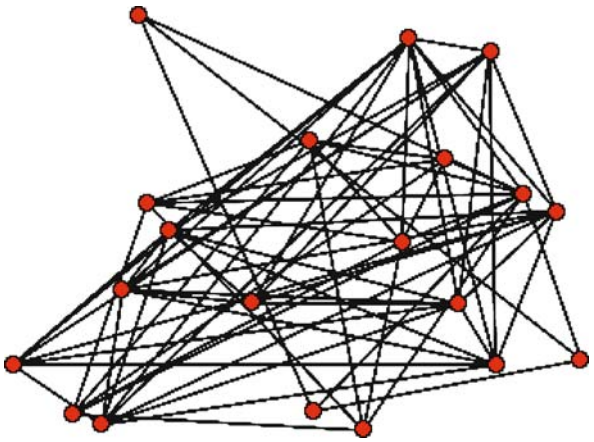
A *path* is a sequence of nodes and edges, beginning with a node that has an edge connecting it to the next node in the sequence. The *length* of a path between two nodes



Names of Participants of Group I	Code Numbers and Dates of Social Events Reported in *Old City Herald*													
	(1) 6/27	(2) 3/2	(3) 4/12	(4) 9/26	(5) 2/25	(6) 5/19	(7) 3/15	(8) 9/16	(9) 4/8	(10) 6/10	(11) 2/23	(12) 4/7	(13) 11/21	(14) 8/3
1. Mrs. Evelyn Jefferson	×	×	×	×	×	×		×	×					
2. Miss Laura Mandeville	×	×	×		×	×	×	×						
3. Miss Theresa Anderson		×	×	×	×	×	×	×	×					
4. Miss Brenda Rogers	×		×	×	×	×	×	×						
5. Miss Charlotte McDowd			×	×	×		×							
6. Miss Frances Anderson			×		×	×		×						
7. Miss Eleanor Nye					×	×	×	×						
8. Miss Pearl Oglethorpe						×		×	×					
9. Miss Ruth DeSand					×		×	×	×					
10. Miss Verne Sanderson							×	×	×			×		
11. Miss Myra Liddell								×	×	×		×		
12. Miss Katherine Rogers								×	×	×		×	×	×
13. Mrs. Sylvia Avondale							×	×	×	×		×	×	×
14. Mrs. Nora Fayette						×	×		×	×	×	×	×	×
15. Mrs. Helen Lloyd							×	×		×	×	×		
16. Mrs. Dorothy Murchison								×	×					
17. Mrs. Olivia Carleton									×		×			
18. Mrs. Flora Price									×		×			

Fig. 3.—Frequency of interparticipation of a group of women in Old City, 1936—Group I.

Social Network Visualization, Methods of, Figure 10
The Davis, Gardner and Gardner data on women's attendance at social events



Social Network Visualization, Methods of, Figure 11
Links in the network of a homeless woman I

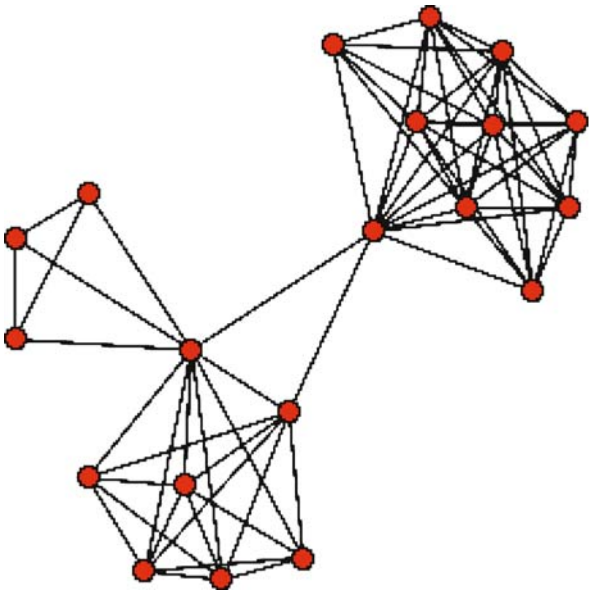
is the number of edges it contains. And the shortest path connecting two nodes is called the *geodesic*.

Any two nodes in *N* are *connected* if there is a path from one to the other. And a whole graph *G* is connected if every pair of nodes in *N* is connected. If a path begins and ends at the same node, that path is a *cycle*. Finally, a graph is a *tree* if it is both connected and it contains no cycles.

The image in Fig. 11 is a graph. It is based on data recorded by J. Clyde Mitchell [14] on the social ties among the 19 individuals involved in the personal network of a homeless woman in Britain. I used a program called NetDraw to place the nodes representing individuals in Fig. 11 in random positions. That calls attention to the importance of the locations of points in graphic displays. Given the locations of the points in Fig. 11 it is very difficult for the viewer to see anything interesting in the patterning of this woman's network.

Compare the image in Fig. 11 with that in Fig. 12. Figure 12 was also produced using NetDraw, but this time the points were placed using a *spring embedder* [16]. A spring embedder is a computer algorithm that, in effect, places a spring of unit length between every pair of adjacent nodes and a much longer spring between nodes that are not adjacent. It starts with a random placement of nodes, then the whole apparatus is set in motion and the various springs push and pull until they reach an equilibrium.

The advantage of using a spring embedder is that it does not require the investigator to make ad hoc judgments in locating nodes in a graph. It uses a standard computer algorithm to place the nodes automatically. Thus, every user will get the same result. There are several different spring embedding algorithms. And they are all examples of a more general class of computer algorithms that search for optimal locations for nodes in relatively few dimensions. This general class of search algorithms is called *multidimensional scaling* [17].



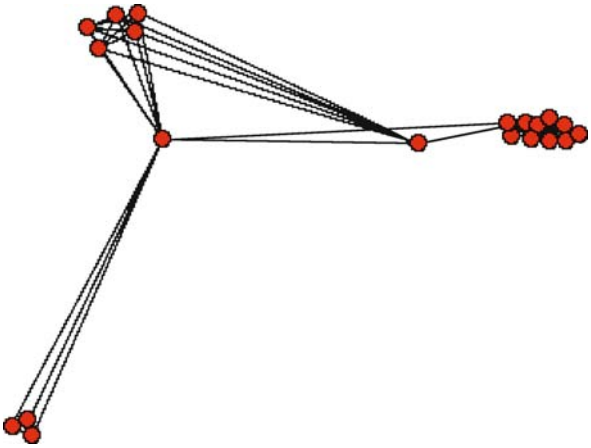
Social Network Visualization, Methods of, Figure 12
Links in the network of a homeless woman II

An alternative method for placing nodes automatically is grounded in algebra. It is called *singular value decomposition* [18]. Singular value decomposition is not search based. Instead, it uses matrix operations to produce a linear transformation of the data, and thus to position the nodes in one, two or three or more dimensions. There is no guarantee that it will always be effective, but often singular value decomposition provides very good placements of the nodes in few enough dimensions that visualization is possible [19]. A NetDraw image based on singular value decomposition of Mitchell's data is shown in Fig. 13.

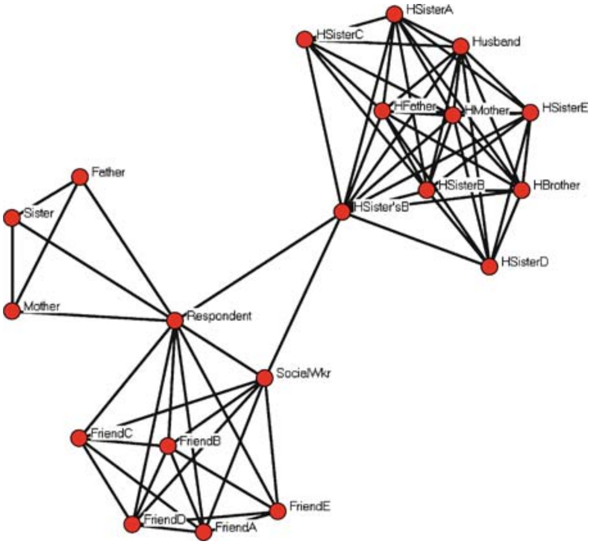
The images in Figs. 12 and 13 both show that the whole network is organized into three densely connected groups that are only loosely linked to one another. That is interesting, but it does not tell us anything about the bases for the groupings. By adding a little information, and continuing to use NetDraw, we can transform the graph of Fig. 12 into a *node labeled* graph, see Fig. 14.

Given the labels, we can identify the homeless woman, the "respondent." We can also see how her network is split up. One division includes her original family, another her friends along with her social worker, and the third contains her estranged husband and his family, her in-laws.

Mitchell's report, however, included even more details. It included estimates of the strength of the tie linking each



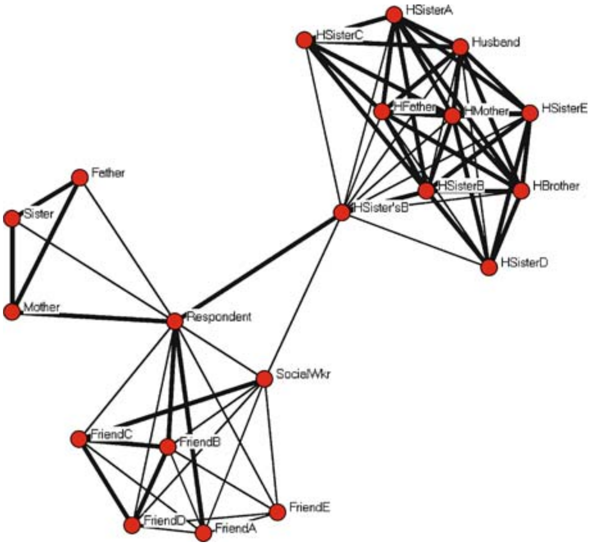
Social Network Visualization, Methods of, Figure 13
Links in the network of a homeless woman III



Social Network Visualization, Methods of, Figure 14
Links in the network of a homeless woman IV

of the pairs of individuals. He classified each tie as either strong or weak. We can embody this additional information in our NetDraw image by an adding another component to our graph. Figure 15, then, was produced using the spring embedder, it is node labeled, and, in addition, it is an *edge labeled* graph.

In Fig. 15 strong ties are indicated by wide edges. By examining their patterning, we learn that the individuals within each family are linked together mostly by strong ties, while the homeless woman's friends have fewer strong ties linking them together. This result is not surprising, but it does provide additional insight about the structural position of the woman in question. Clearly, it would be easier for either family to achieve consensus and provide support than it would be for the respondent's loosely connected collection of friends [20].



Social Network Visualization, Methods of, Figure 15
Links in the network of a homeless woman V

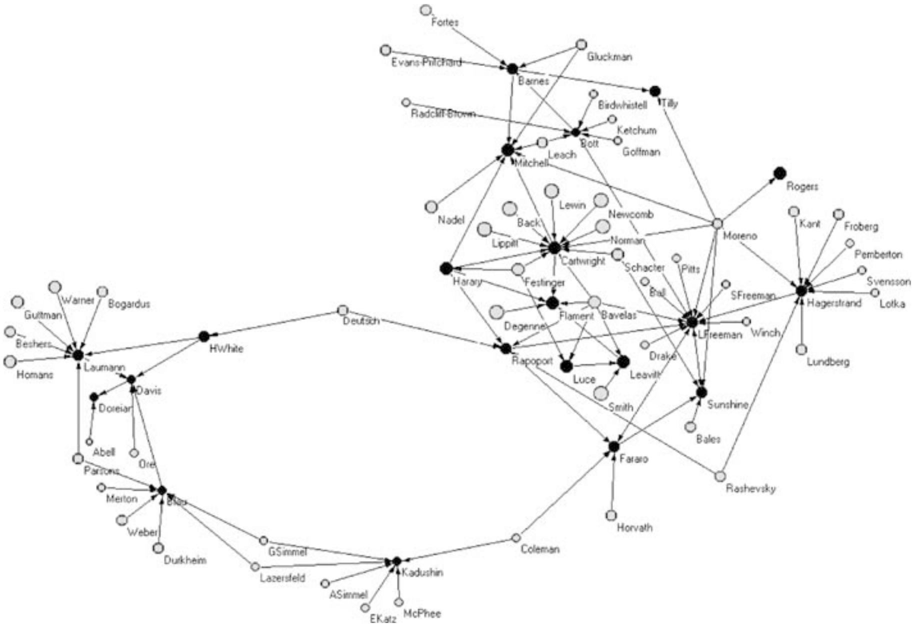
It should be clear, then, that the placement of nodes and the labeling of both nodes and edges are critical for the ability of a graph to communicate important information. Good images can provide investigators with new insights about the structural properties of the social networks they are studying. And they can, of course, help to communicate the results of social network research to outsiders.

Images Based on One Mode Directed Relations

It was obvious from the outset that these simple graphs would not permit many kinds of displays of interest to social network analysts. Even Moreno and Jennings saw the need to display the direction of choice in their sociograms. The direction of connections can be expressed using *directed graphs* or *digraphs*.

A digraph $D = \langle N, A \rangle$ where N is a finite collection of nodes and A is a set of pairs shown as *directed lines* or *arrows*. When an arrow is directed from node a to node b in a digraph, then a is the *tail* of the arrow and b is the *head*; a is the immediate *predecessor* of b and b is the immediate *successor* of a. The *outdegree* of a node is the number of arrows for which it is the tail and its *indegree* is its number for which it is the head.

In any study that involves social links that are not symmetric, digraphs provide a natural representation. Con-

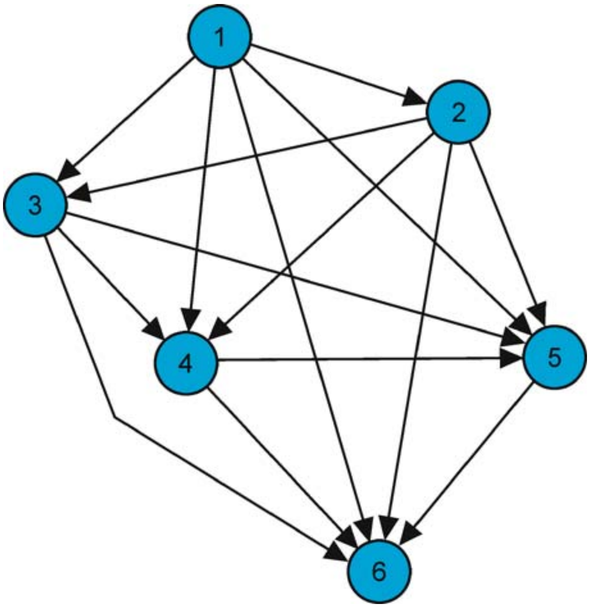


Social Network Visualization, Methods of, Figure 16
Influences on some founders of social network analysis

sider Fig. 16 that was produced by a program called Pajek [21]. In preparing a book on the development of social network analysis, I interviewed a number of the founders. Each was asked to name others who had influenced them to think in network terms. The result is a data set that obviously lacks symmetry.

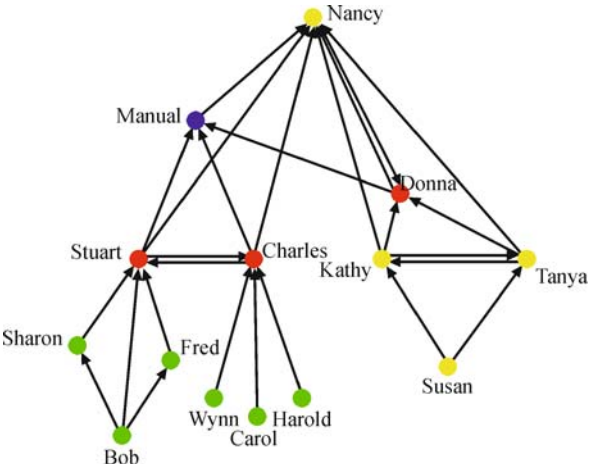
My interest, however, was with clusters, or blocs, of influentials and nominees. So I placed the nodes using a spring embedder designed by Kamada and Kawai [22]. The resulting figure shows that there seem to be two fairly well defined subgroups, one on the left and one on the right. The two groups are relatively dense but they are only loosely connected together. The people on the left are almost entirely sociologists and those on the right are mostly from other fields. And from the patterning one can suspect that there was some kind of split between these two groups.

For some kinds of data the search for clusters or groups is not appropriate. For example, when we are dealing with data that should embody some sort of ordering, digraph representations are particularly important. To illustrate how digraphs can be used to display ordering, consider the data collected by Forkman and Haskell [23]. They studied several communities, each made up of six domestic hens. In five of these communities the hens formed strict pecking orders in which the top hen pecked all the others; the second pecked all but the top, and so on. Figure 17 shows a visone [24] image of the data from one of those five communities. There the nodes are arranged, top down, in terms of their outdegrees and the pecking order is obvious.



Social Network Visualization, Methods of, Figure 17
Dominance among six hens

Often data approach, but do not achieve, a strict order. David Krackhardt [25], for example, collected data on who sought advice from whom among 14 employees in the in-



Social Network Visualization, Methods of, Figure 18
Visone image of advice seeking (from [26])

ternal auditing staff of a large company. Krackhardt's data could not be drawn with all the arrows pointing in one direction. So, in Fig. 18 he arranged the individuals in such a way that as many arrows as possible were pointing up. The viewer, then, can immediately see there is an important hierarchical element displayed by these data. From the image, it appears that Nancy is at the top of the advice chain and Bob, Wynn, Carol, Harold and Susan are at the bottom.

There is, however, an important limitation in this figure. Nancy seeks advice from Donna, Donna seeks advice from Manuel and Manuel seeks advice from Nancy. Thus, these three form a directed cycle of advice seeking. Given such a circular arrangement, no possible hierarchy among these three individuals can be established. Any order in which they were arranged would be misleading. In addition, Stuart and Charles cannot be ordered because they chose each other. The same is true for Kathy and Tanya.

The apparent ordering of nodes in Krackhardt's image was imposed by human judgment. There are computer algorithms that can automatically arrange the nodes into a hierarchical form [26]. They are, however, not as well grounded or reliable as multidimensional scaling and singular value decomposition.

Images Based on Two Mode Relations

Any time we deal with a relation that can link more than two social actors, we cannot use graphs or directed graphs. Both graphs and directed graphs can deal only with links between pairs. Two mode data, however, allow for relations that link three or more actors. So, whenever we have two mode data, like that collected by Hobson [8] or Davis, Gardner and Gardner [13] we need another way to construct images.

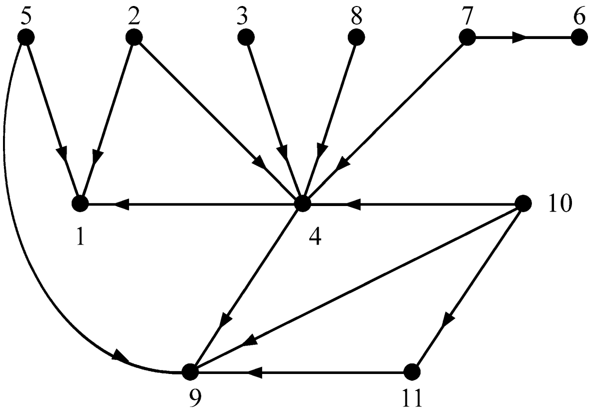
There are several ways to construct images of two mode data. I will consider three of them in the present section, hypergraphs, bipartite graphs and lattices. Then, in the next section, I will discuss the use of matrix representations for both one mode and two mode data.

Hobson [8] collected two mode data on corporations and their directors. He produced the image shown in Fig. 4 showing corporate interlocks as overlapping areas. Mathematically, images like Hobson's are *hypergraphs*. A hypergraph, $F = \langle N, H \rangle$, consists of a set of nodes N and a collection of *hyperedges*, H. While an edge in an ordinary graph connects two nodes, a hyperedge in a hypergraph may link any arbitrary subset of the nodes in N. Pictorially, hyperedges are represented as boundaries enclosing sets of nodes.

The use of hypergraphs was demonstrated in a recent report by Estrada and Rodríguez-Velázquez [27]. They began with one mode data that showed the patterning of predation among the members of eleven species in a Malaysian rain forest. Their graph, showing who preys on whom, is shown in Fig. 19.

Figure 19 shows which species preys on which other species. But if the investigator is interested, as those who study food webs often are, in defining ecological niches in terms of co-predation, Fig. 19 makes the overall pattern less than obvious. As an alternative we can build a hypergraph.

The matrix shown in Fig. 20 is based on the data in Fig. 19. It was built by considering each of the species in turn as prey. Then all of the species that share each given prey are pooled together. Species 1, 6 and 9 have no prey

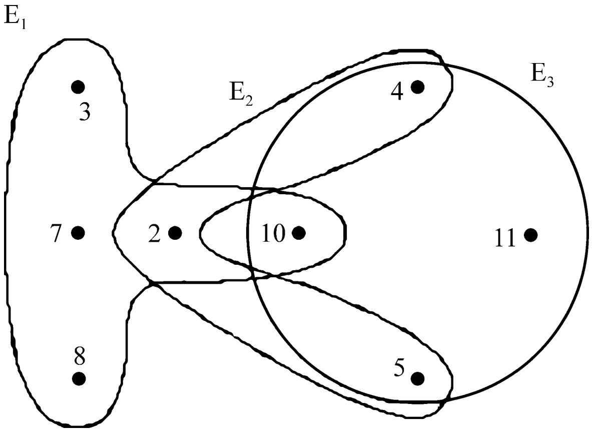


Social Network Visualization, Methods of, Figure 19
Who preys on whom in a Malaysian Rain Forest

Prey on

	4	1	9
2	1	1	0
3	1	0	0
4	0	1	1
5	0	1	1
7	1	0	0
8	1	0	0
10	1	0	1
11	0	0	1

Social Network Visualization, Methods of, Figure 20
Two mode matrix of co-predation



Social Network Visualization, Methods of, Figure 21
Hypergraph of co-predation

in the set. And species 4, 1 and 9 are the targets of co-predation. So the new matrix is two mode. It has the three targets of co-predation as columns and the eight predators as rows.

That matrix is captured visually by the hypergraph in Fig. 19. It immediately reveals that there are three niches. The one labeled E_1 includes all the species who preyed on species 4, E_2 those who preyed on species 1 and E_3 those who preyed on 9. Thus, each edge in Fig. 21 encloses a collection of species that compete directly for at least one prey.

There are, however, other ways to picture two mode data. In a more recent study of corporate interlocks, Joel Levine [29] reported data on the board memberships of seven major American corporations. Those corporations turned out to have ten directors who appeared on two or more of their boards. Levine presented his interlock data using a *bipartite graph*. A bipartite graph, $B = \langle N, E \rangle$ where N is partitioned into two disjoint subsets, N_1 and N_2, and no edge in E has both end points in the same subset. He used singular valued decomposition to place the nodes representing both corporations and board members and produced a bipartite image similar to the one displayed in Fig. 22. I prepared that figure using NetDraw. There, the corporations are shown as red circles and the board members are blue squares. Thus, both the colors and the shapes of the nodes stress the bipartite nature of the graph.

There is still another form of graphic display, one that reveals even more structural information about a two mode data set. It is based on an algebraic procedure called *Galois lattice analysis* or *formal concept analysis* [30,31,32,33]. A Galois or formal concept lattice is defined on an object by property, matrix. Let O be a set of objects and A be a set of attributes. The binary matrix $O \times A$ indicates which objects possess which attributes.

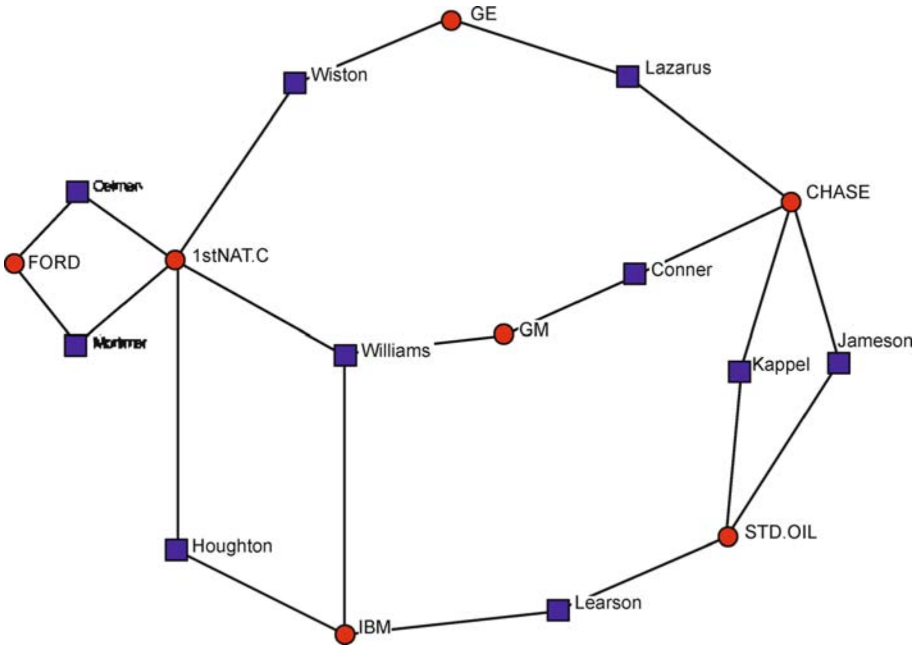
We can define a pair $\langle O_i, A_i \rangle$ such that O_i is a subset of O and A_i is a subset of A and every object in O_i has every attribute in A_i. Moreover, both O and A must be maximal. Thus, for every attribute in A that is not in A_i, there is an object in O_i that does not have that attribute. And for every object in O that is not in O_i, there is an attribute in A_i that the object lacks.

These pairs are dual and they can be partially ordered by inclusion. Given two pairs $\langle O_i, A_i \rangle$ and $\langle O_j, A_j \rangle$ we say that $\langle O_i, A_i \rangle$ is less than $\langle O_j, A_j \rangle$ when O_i is a subset of O_j or, equivalently, when A_j is a subset of A_i. Since all these pairs have unique least upper bounds and greatest lower bounds they form a dual (Galois) lattice.

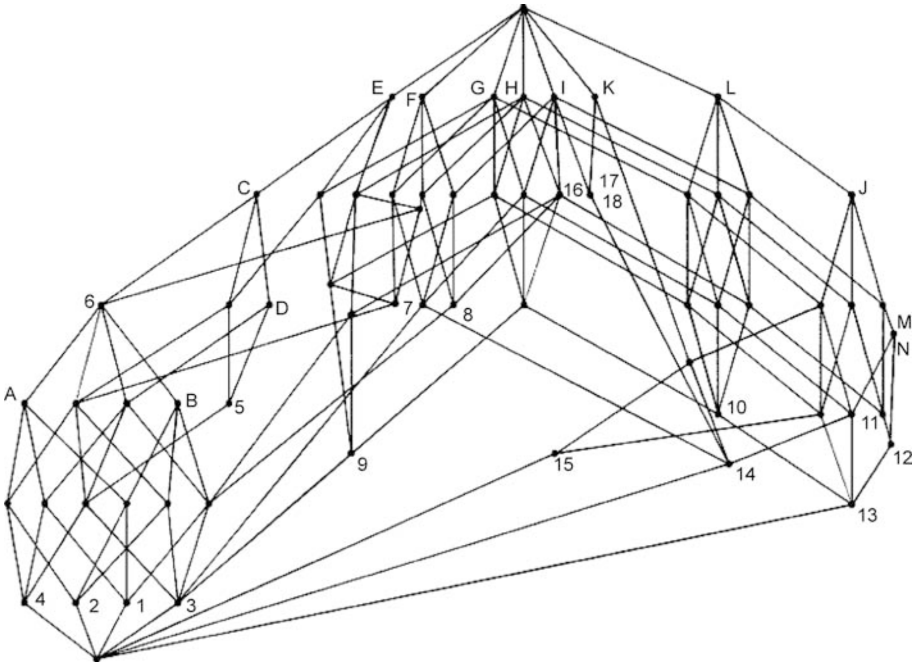
I will illustrate by considering again the woman by event data collected by Davis, Gardner and Gardner [13]. Let the women (1 through 18) be the objects and the events (A through N) be the attributes. The data, arranged into a Galois lattice by a program called GLAD, are shown in Figure into a lattice in Fig. 23.

The lattice displays the same three classes of events that define the same two groups of women that we saw in Fig. 10. But, in addition to the classes of events and groups of women, we can now see the containment structures of both events and women. To begin with, by following lines up from the bottom we can see which women attended which events. When we get to the top we hit the set of all events and, at the same time, because no woman attended all 14 events, it is also the null set of women.

The uppermost events (E–L) involved the largest sets of women. Other events are contained in the lower inter-



Social Network Visualization, Methods of, Figure 22
A NetDraw Image of Levine's interlock data as a bipartite

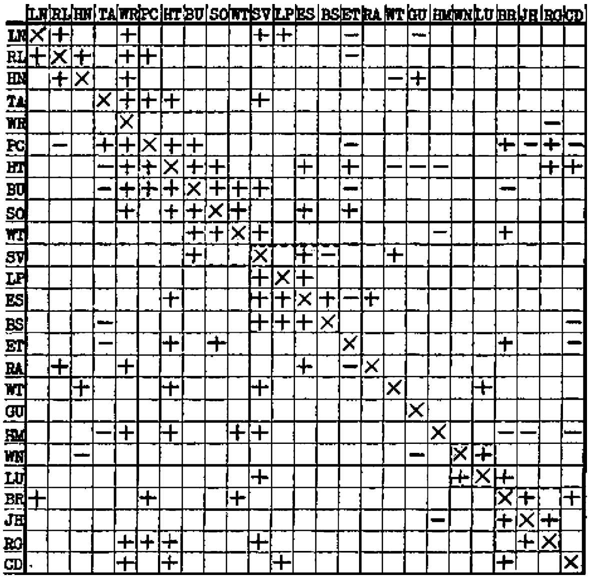


Social Network Visualization, Methods of, Figure 23
The Davis, Gardner and Gardner data as a Galois lattice

sections of these events. Event C, for example, is contained in E; everyone who attended C was present at E. And, at the next lower level, B and D are both contained in C. The events, then, can be seen as varying in their "openness".

At the same time, the figure shows the upward containment structure of the women in terms of their patterns of attendance. Because no event attracted all 18 women, the lowest point represents the set of all women as well as the null set of events. Then, the lowest set of women (1, 2, 3, 4, 13, 14 and 15) are the "core" attendees, so to speak. The next level contains woman 9 who never attended unless woman 3 was also present, and woman 5 whose attendance depended on that of women 4 and 3. Women 6, 7, 8, 10, 11, 12, 17 and 18 are also at this second level. In some sense, these are all secondary or peripheral participants in these events. And, finally, woman 16 turns out to be a third level participant; she was extremely peripheral. Woman 16 attended events only when secondary attendees 8–12 and core attendees 1, 3 and 13 were all present. All in all, then, the image of the Galois lattice reveals a great deal about the internal structure of attendance.

In this section I have shown three ways of visualizing two mode data. All three of them, however, share one important limitation. That limitation stems from the fact that all three of them can only be used for very small data sets. As the number of cases grows, they all produce images that become increasingly difficult to read.



	LN	RL	HN	TA	WR	PC	HT	BU	SO	WT	SV	LP	ES	BS	ET	RA	WT	GU	HM	WN	LU	BR	JH	RG	GD
LN	X	+			+						+	+			−			−							
RL	+	X	+		+	+									−										
HN		+	X		+												−	+							
TA				X	+	+	+				+														
WR					X																			−	
PC		−		+	+	X	+	+							−							+	−	+	−
HT				−	+	+	X	+	+				+		+		−	−	−					+	+
BU				−	+	+	+	X	+	+	+				−							−			
SO					+		+	+	X	+			+		+										
WT								+	+	X	+								−			+			
SV								+			X		+	−			+								
LP											+	X	+												
ES							+				+	+	X	+	−	+									
BS				−							+	+	+	X											−
ET				−			+		+						X							+			−
RA		+			+								+		−	X									
WT			+				+				+						X				+				
GU																		X							
HM				−	+		+			+	+								X			−	−		−
WN			−															−		X	+				
LU											+									+	X	+			
BR	+					+				+												X	+		+
JH																			−			+	X	+	
RG					+	+	+				+												+	X	
GD					+		+					+										+			X

Social Network Visualization, Methods of, Figure 24
The Forsyth and Katz image of sociometric choices

Images Based on One or Two Mode Data Matrices

When Davis, Davis and Gardner [13] first used matrix permutation, they did so without calling attention to the process. But since that first use a number of contributors have suggested procedures explicitly designed to rearrange the rows and columns of matrices. As time has passed, the overall tendency has been to come up with more effective procedures. And, with the introduction of computers, it has become possible to manipulate ever larger matrices. Presently, there is no end in sight.

Matrix permutations, moreover can be used with either one mode or two mode data. Five years after Davis, Gardner and Gardner introduced matrix permutation in their two mode data set, Elaine Forsyth and Leo Katz [34] explicitly proposed permuting matrices as a way to uncover and display social groups in a one mode data set. They illustrated using data from one of Moreno's [10] sociometric studies. The young women in a residence hall had each been asked to name others in their hall for whom they had positive feelings and those for whom their feelings were negative. Positive choices were recorded using plus signs and negative choices were recorded as minus signs.

Forsyth and Katz adopted a brute-force procedure that involved rearranging rows and columns and redrawing the image again and again until as many of the plus signs fell as close to the principle diagonal as possible. At that point, cohesive groups become visible as clusters of plus signs around the diagonal. Their result is shown in Fig. 24.

Obviously, the Forsyth and Katz procedure was extremely cumbersome. But Beum and Brundage [35] soon came up with a systematic iterative procedure for finding groups by rearranging the rows and columns of a one mode matrix. And, by the late 1950s, when computers emerged on the scene, Coleman and MacRae [36] developed a series of Univac programs at the Operations Analysis Laboratory at the University of Chicago that were designed to uncover the groups in large networks.

An entirely different kind matrix permutation procedure was proposed by Harrison White and his students. They introduced the idea of blockmodeling [37]. In so doing, they provided a theoretical basis for reordering network data matrices, and they developed a number of algorithms for doing so.

The aim of this new thrust was to reorder the matrix in such a way that it could be partitioned to reveal two or more collections of social actors who were *not* linked by some social relation of interest. So, instead of arraying actors along the diagonal of a matrix, White et al. sought permutations that would define zero blocks – sets of actors between which there were no social links. They used their approach to examine a great many network data sets. One example is shown in Fig. 25.

The data in Fig. 25 were collected by Sampson [38] in his study of a monastery. Sampson asked each of a collection of 18 novices to report their relationships with each of the others. Figure 25 shows an 18 by 18 matrix of their responses to a question asking the novices about which oth-

	1	2	3	4	5	6	7	8	9	1	11	12	13	14	15	16	17	18
1										2						1		3
2				3									2					
3		1		3		2												
4	2	3												1				
5			1														3	2
6							3										2	1
7			2	2		3		2									1	
8												1	2			3		
9												1					3	2
1																		
11							3										1	2
12																		
13		3				2						1						
14				2									3				1	1
15			1	3									2					
16			1	3							2							
17				2		3		1										
18				3		1					2							

Social Network Visualization, Methods of, Figure 25
Sampson's data on who was a negative influence on whom

ers had negative influences on them. A response of 3 indicated a first choice. A 2 was a second choice and a 1 was a third choice. White et al. reasoned that only first and second choices represented strong responses, so they ignored the third choices and treated the entries of 1 as if they did not exist.

One of the several procedures procedure White et al. introduced was called CONCOR. CONCOR is a recursive procedure that begins by calculating correlations between the rows (or columns) of a network data matrix. Then correlations are calculated between the rows of the resulting correlation matrix. That procedure continues until it produces a matrix of correlations that uniformly displays values of +1 and −1. Those positive and negative values are used to partition the individuals into two subsets. The CONCOR procedure can be repeated using the data contained within each of the partitions. Thus, the original matrix can thus be refined to any desired degree.

White and his students used CONCOR on the data shown in Fig. 25 in an attempt to uncover blocks that contained only 0s. They could then use these zero blocks to reduce the complexity of the data matrix. That matrix produced three zero blocks. They are shown in Fig. 26.

In the reduced model in Fig. 27 each cell represents one of the blocks in Fig. 26. Thus, the 18 by 18 matrix is reduced to a 3 by 3 array. The reduction is consistent with Sampson's original ethnographic description of subgroups among the novices. Moreover, its pattern of zero blocks in the principle diagonal indicates that no block member saw any fellow block member as having a negative influence. But the members of each block saw at least some of the members of both of the other blocks as negative influences. This makes sense in the light of the ongoing conflict that Sampson described in his report.

Since that time, displays based on matrix permutations have grown in size, complexity and sophistication. One

	10	5	9	6	4	11	8	12	1	2	14	15	7	16	13	3	17	18
10																		
5																1	3	2
9								1									3	2
6													3				2	1
4									2	3	1							
11													3				1	2
8								1						3	2			
12																		
1	2													1				3
2					3										2			
14					2										3		1	1
15					3										2	1		
7				3	2		2									2	1	
16					3	2										1		
13				2				1		3								
3				2	3					1								
17				3	2		1											
18				1	3	2												

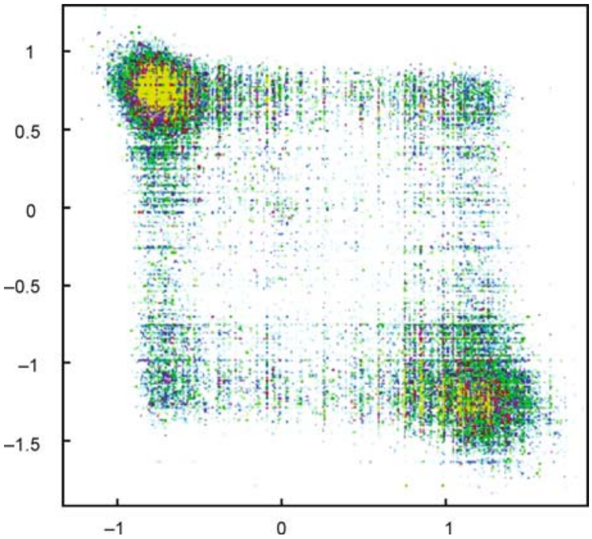
Social Network Visualization, Methods of, Figure 26
White, Boorman and Breiger's partitioning of the negative influence data matrix from Sampson

0	1	1
1	0	1
1	1	0

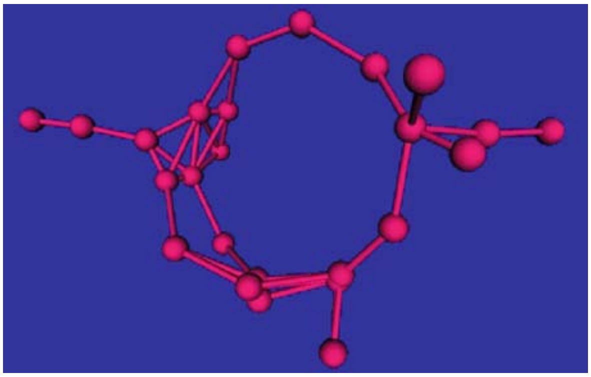
Social Network Visualization, Methods of, Figure 27
The CONCOR reduction of the Sampson matrix

particularly striking example was produced by Richards and Seary [39]. Their data were drawn from a study of participants in a needle exchange program in Baltimore, Maryland [40]. Richards and Seary examined data on 4259 individuals who picked up and returned needles at each of four exchange sites over a 30 month period. Each cell in the matrix is a record of the number of needles picked up by the individual in that row and returned by the individual in that column. About a third of all needles fall in the principle diagonal of the matrix.

Richards and Seary used the data from the largest weak component in the data set. That component involved 100 000 needle exchanges among 36 000 individuals. Richards and Seary used their program MultiNet [41] scaled the data using a form of singular value decomposition called *correspondence analysis* [18]. They used the coordinates provided by the first Eigenvector to reorder the rows and columns of the matrix. They then colored the entries in terms of frequencies. The color scale is logarithmic: gray is 1 needle, blue 2–3, green 4–7, red 8–15, ma-



Social Network Visualization, Methods of, Figure 28
The largest component in the Baltimore needle exchange data



Social Network Visualization, Methods of, Figure 29
VRML image of friendship among teens in a Dublin suburb

genta 16–31, yellow 32 and above. Their image is shown in Fig. 28.

Figure 28 dramatically illustrates the utility of images based on matrix permutation. It shows that there was not a single community of needle users in Baltimore. Instead, there were two distinct communities of individuals who regularly obtain, return and exchange needles with one another. These two relatively large communities were centered around two of the four needle exchange sites.

Future Directions

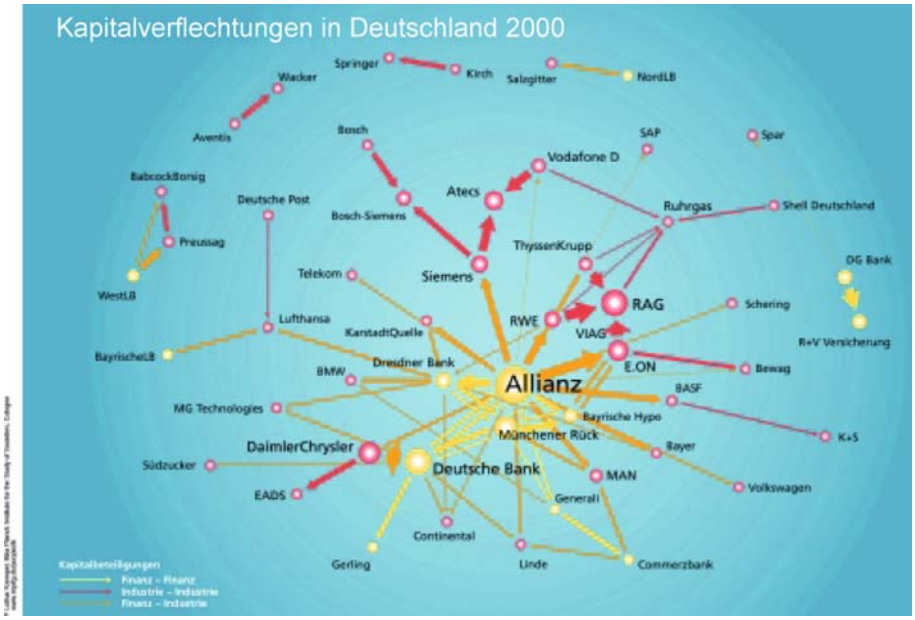
Overall, the long range the trend in visualizing social networks has been to rely on computers to do more and more of the job. First, computers used a version of singular value decomposition to locate nodes in two dimensional images [42]. Then, soon thereafter, Coleman and MacRae [35] programmed a computer both to permute rows and columns of a matrix and to print out an image of the result. And, in the early 1970s, Alba [43] wrote a program that performed calculations to place nodes and then went on to draw node and edge images of the results [44].

Since the 1970s, then, network analysts have increasingly used computers both for calculations and to draw images. And increasingly, multidimensional scaling and singular value decomposition have been used to determine locations for nodes. Moreover, when two dimensions are not enough to display network structure, three dimensional images are being produced.

When microcomputers became available it quickly became possible to produce images that gave the appearance of being three dimensional. Moreover, with the advent of color screens, color images began to be produced. Figure 29 represents data collected by Kirke [45] on social links among teenagers in a suburb of Dublin. Nodes were first located in three dimensions using multidimensional scaling. And then the virtual reality modeling language (VRML) was used to produce the appearance of three dimensions. Figure 27 was produced as a cover design for a book [46] and the colors were used simply make the image more attractive.

Colors can, however, be used to enhance the ability of an image to communicate important information. In Fig. 30 Höpner and Krempel [47] used a spring embedder and Krempel's own programs to arrange the nodes in two dimensions. The nodes represent the 100 largest German corporations in the year 2000. They used color to label both nodes and directed lines. In their image each company is represented as a node and an arrow pointing from one node to another means that the first node holds shares in the second. The size of a node indicates the number of connections to other nodes it has. Financial companies are shown as yellow nodes and industrial companies are red. Links between financial companies are yellow, those between industrial companies are red and links between financial and industrial companies are orange. Bu using color, then, his directed graph reveals a great deal of information about the organization of German industry and finance.

The image in Fig. 31 was made using a program called MAGE [48]. MAGE was written by Richardson and Richardson [49]. It is designed as a display on computer screens, and it allows the viewer to move into the picture as well as to spin and rotate it. It is useful, then, for exploring the patterning of structural data in three apparent dimensions.



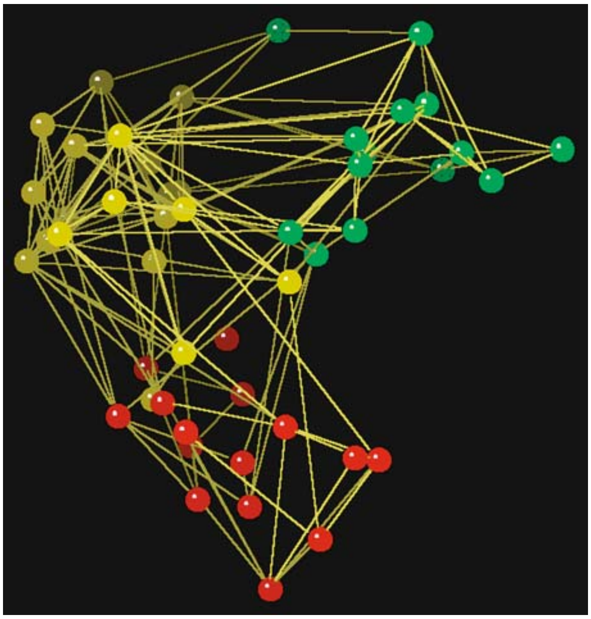
Social Network Visualization, Methods of, Figure 30
Links among German corporations in 2000

In Fig. 31 students in a university professional school program reported who their friends in their class were. Nodes were placed using a multidimensional scaling program and then they were colored according to their program in the school. It turned out that most of their friendship choices linked those that shared a program.

Richards and Seary's [50] program, Multinet, produces a wide range of graphic images. Included are images that actually can be viewed in three dimensions using *anaglyphic* glasses in which one lens is red and one is blue. Obviously, I cannot illustrate their program here, but anyone who wants to see real 3D images should explore Multinet.

The most recent development in visualizing social networks involves the production of animated graphics. As more and more process data are collected and as more process models are constructed, animated images are a natural development. A group at Stanford University has written a Java program, SoNIA [51], that makes it quite simple to produce animated node and edge and node and directed line images [52,53]. These images allow users to explore the changing structural forms generated by process data.

Overall then, in the period between Moreno's hand drawn ad hoc images and the latest animations of dynamic network processes, there has been a dramatic growth in our ability to visualize social network structure. The major contribution has come from computers. Today we can use a wide variety of readily available computer programs to both design images and to produce screen images and/or printed output.



Social Network Visualization, Methods of, Figure 31
MAGE image of friendship among classmates

But, as the job of producing images becomes easier, we must be careful not to lose our sense of why we are producing then in the first place. From the very beginning, the important point has always been that the visual images of social networks are not produced simply to be decorative. In every case, the early images were drawn in order to dramatize some feature of social structure. Moreno produced Fig. 5 to illustrate the importance of considering the number of connections in evaluating the structural position of an individual. In Fig. 6 the number of negative ties received by one of the running backs showed, as Moreno (p. 213 in [10]) put it, "It is easy to see that when 5/RB is running with the ball he is not apt to get the maximum of cooperation in interference and blocking."

Figure 7 was a pictorial statement by Warner and Lunt that when cohesive subgroups overlap, they should not be expected to bridge wide differences in social class. Figure 8, from Davis, Gardner and Gardner, demonstrated that the Warner–Lunt hypothesis was supported by data with respect to both social class and age. And, finally, Fig. 9 illustrated the presence of cohesive groups and of the variation of different individuals in their involvement in those groups. In every case each of these early authors had a point to make, and in every case the image helped to make that point. That is the key to the effective use of visual materials in social network analysis.

In future we can expect to see continued development of computer programs designed to aid in visualizing social networks. We can look forward to continued refinement of algorithms for displaying group structure that are based on multidimensional scaling, particularly spring embedding. We can anticipate better algorithms for displaying hierarchies and approximate hierarchies. We can expect to have more powerful programs for animation. And, at the same time, we can expect to be able to produce higher quality and more refined visual displays of all sorts.

Bibliography

Primary Literature

1. Freeman LC (2004) The development of social network analysis: A study in the sociology of science. Empirical Press, Vancouver, BC
2. Schweitzer VS (1998) Words of power. Coffee Times, Fall
3. Cayley A (1857) On the theory of the analytical forms called trees. Philos Mag 13:19–30
4. Biggs NL, Lloyd EK, Wilson RJ (1977) Graph theory 1736–1936. Oxford University Press, Oxford
5. Morgan LH (1871/1997) Systems of Consanguinity and Affinity in the Human Family. University of Nebraska Press, Lincoln
6. Macfarlane A (1883) Analysis of relationships of consanguinity and affinity. J R Anthropol Inst Great Britain Ireland 12:46–63
7. Appendix to Macfarlane A (1883) Analysis of relationships of consanguinity and affinity. J R Anthropol Inst Great Britain Ireland 12:46–63
8. Hobson JA (1884) The evolution of modern capitalism; A study of machine production. Allen & Unwin, Macmillan, London, New York
9. Moreno JL (1932) Application of the group method to classification. National Committee on Prisons and Prison Labor, New York
10. Moreno JL (1934) Who shall survive? Nervous and Mental Disease Publishing Company, Washington
11. Roethlisberger FJ, Dickson WJ (1939) Management and the worker. Harvard University Press, Cambridge
12. Warner WL, Lunt PS (1941) The social life of a modern community. Yale University Press, New Haven
13. Davis A, Gardner B, Gardner MR (1941) Deep south. University of Chicago Press, Chicago
14. Mitchell JC (1994) Situational analysis and network analysis. Connections 17:16–22
15. Available at http://www.analytictech.com/downloadnd.htm
16. Eades P (1984) A heuristic for graph drawing. Congressus Nutnerantiunt 42:149–160
17. Krempel L (1999) Visualizing networks with spring embedder: Two-mode and valued data. In: Proceedings of the section of statistical graphics. American Statistical Association, Alexandria, pp 36–45
18. Weller SC, Romney AK (1990) Metric scaling: Correspondence analysis. Sage Publications, Newbury Park
19. Freeman LC (2005) Graphical techniques for exploring social network data. In: Carrington PJ, Scott J, Wasserman S (eds) Models and methods in social network analysis. Cambridge University Press, Cambridge, pp 248–269
20. Bott E (1957) Family and social network. Tavistock, London
21. Available at http://vlado.fmf.uni-lj.si/pub/networks/pajek/
22. Kamada T, Kawai S (1989) A general framework for visualizing abstract objects and relations. ACM Trans Graph 10:1–39
23. Forkman B, Haskell MJ (2004) The maintenance of stable dominance hierarchies and the pattern of aggression: Support for the suppression hypothesis. Ethology 110:737–744
24. Available at http://visone.info/download/
25. Krackhardt D (1996) Social networks and the liability of newness for managers. In: Cooper CL, Rousseau DM (eds) Trends in organizational behavior, vol 3. Wiley, New York, pp 159–173
26. Brandes U, Raab J, Wagner D (2001) Exploratory network visualisation: Simultaneous display of actor status and connections. J Soc Struct 2:4
27. Estrada E, JA Rodríguez-Velázquez. (2005) Complex networks as hypergraphs. Physica A 364:581–594
28. Available on request from its author, Vincent Duquenne (v.duquenne@wanadoo.fr)
29. Levine JH (1979) Joint-space analysis of 'pick-any' data: Analysis of choices from an unconstrained set of alternatives. Psychometrika 44:85–92
30. Wille R (1982) Restructuring lattice theory: An approach based on hierarchies of concepts. In: Rival I (ed) Ordered Sets. Reidel, Dordrecht–Boston, pp 445–470
31. Wille R (1984) Line diagrams of hierarchical concept systems. Int Classif 11:77–86
32. Duquenne V (1987) Contextual implications between attributes and some representation properties for finite lattices. In: Ganter B, Willie R, Wolff K (eds) Beiträge zur Begriffsanalyse. B. I. Wissenschaftsverlag, Berlin, pp 213–240

33. Freeman LC, White DR (1993) Using Galois lattices to represent network data. In: Marsden PV (ed) Sociological methodology. Blackwell, Oxford, pp 127–146
34. Forsyth E, Katz L (1946) A matrix approach to the analysis of sociometric data: Preliminary report. Sociometry 9:340–347
35. Beum CO, Brundage EG (1950) A method for analyzing the sociomatrix. Sociometry 13:141–145
36. Coleman JS, MacRae D (1960) Electronic processing of sociometric data for groups up to 1000 in size. Am Sociol Rev 25:722–727
37. White HC, Boorman SA, Breiger RL (1976) Social structure from multiple networks: I. Blockmodels of roles and positions. Am J Sociol 81:730–779
38. Sampson SF (1969) A noviate in a period of change: An experimental and case study of relationships. Unpublished Ph D dissertation. Department of Sociology, Cornell University
39. Richards WD, Seary AJ (2000) Cover illustration. Connections 23:1
40. Valente TW, Foreman RK, Junge B, Vlahov D (1998) Satellite exchange in the Baltimore needle exchange program. Public Health Reports 113:91–96
41. Availble by contacting Bill Richards (richards@sfu.ca)
42. Bock RD, Husain SZ (1952) Factors of the tele: A preliminary report. Sociometry 15:206–219
43. Alba R (1972) SOCK. Behav Sci 17:326–327
44. Kadushin C (1974) The american intellectual elite. Little, Brown, Boston
45. Kirke DM (1996) Collecting peer data and delineating peer networks in a complete network. Social Netw 18:333–346
46. Kirke DM (2006) Teenagers and substance use: Social networks and peer influence. Palgrave Macmillan, Basingstoke, Hampshire and New York
47. Höpner M, Krempel L (2003) The politics of the German company network. Max Planck Institute for the Study of Societies, Working Paper 03/9
48. MAGE is available at http://kinemage.biochem.duke.edu/software/mage.php
49. Richardson DC, Richardson JS (1992) The kinemage – a tool for scientific communication. Protein Sci 1:3–9
50. Richards WD, Seary AJ (2006) Multinet. http://www.sfu.ca/~richards/Multinet/Pages/multinet.htm
51. SoNIA is available at http://www.stanford.edu/group/sonia/
52. Moody J, McFarland DA, Bender-deMoll S (2005) Visualizing network dynamics. Am J Sociol 110:1206–1241
53. Bender-deMoll S, McFarland DA (2006) The art and science of dynamic network visualization. J Social Struct 7:2

Books and Reviews

Brandes U, Kenis P, Raab J, Schneider V, Wagner D (1999) Explorations into the visualization of policy networks. J Theor Polit 11:75–106

Brandes U, Corman SR (2003) Visual unrolling of network evolution and the analysis of dynamic discourse. Information Visualization 2:40–50,

Brandes U, Kenis P, Wagner D (2003) Communicating centrality in policy network drawings. IEEE Transactions on Visualization and Computer Graphics 9:241–253,

Brandes U, Wagner D (2004) Netzwerkvisualisierung. Inform Technol 46:129–134,

Brandes U, Wagner D (2004) Visone - Analysis and visualization of social networks. In: Jünger M, Mutzel P (eds) Graph drawing software. Springer-Verlag, pp 321–340

Brandes U, Kenis P, Raab J (2006) Explanation through network visualization. Methodology 2:16–23

Frank O (2000) Structural plots of multivariate binary data. J Social Struct 1:4

Freeman LC (2000) Visualizing social groups. In: Proceedings of the section on statistical graphics (1999). American Statistical Association, pp 47–54

Freeman LC (2000) Visualizing social networks. J Social Struct 1:1

Freeman LC, Webster CM, Kirke DM (1998) Exploring social structure using dynamic three-dimensional color images. Social Netw 20:109–118

Johnson JC, Krempel L (2004) Network visualization: The 'Bush Team' in Reuters News Ticker 9/11–11/15/01. J Social Struct 5:1

Klovdahl AS (1981) A note on images of networks. Social Netw 3:197–214

Klovdahl AS (1998) A picture is worth...: Interacting visually with complex network data. In: Liebrand WBG (ed) Computer modeling of dynamic social processes. Sage, London

Krempel L (2005) Visualisierung komplexer Strukturen. Grundlagen der Darstellung mehrdimensionaler Netzwerke. Campus, Frankfurt aM

Krempel L, Schnegg M (2005) About the image: Diffusion dynamics in an historical network. Structure and dynamics: eJ Anthropol Rel Sci 1:1, 10

McGrath C, Blythe J, Krackhardt D (1996) Seeing groups in graph layouts. Connections 19:22–29

McGrath C, Blythe J, Krackhardt D (1997) The effects of spatial arrangements on perceptions of graphs. Social Netw 19:223–242

McGrath C, Krackhardt D, Blythe J (2003) Visualizing complexity in networks: Seeing both the forest and the trees. Connections 25:37–47

McGrath C, Blythe J (2004) Do you see what i want you to see? The effects of motion and spatial layout on viewers' perceptions of graph structure. J Social Struct 5:2

Social Organizations with Complexity Theory: A Dramatically Different Lens for the Knowledge Economy

RUSS MARION
Clemson University, Clemson, USA

Article Outline

Glossary

Many of these terms are commonly understood by complexity theorists; these definitions, however, describe their meaning within the context of social organizations.

Adaptive behaviors Perturbations (creative ideas, pressures, etc.) in a system that foster some observable level of phase transition. Change leading to systemic reaction.

Adaptive tension Pressures external or internal to the organization that perturb an organization thus pressuring it toward phase transitions and structural or ideational elaboration.

Administrative leadership behaviors related to such things as strategic planning, policy making, and resource acquisition and distribution.

Agent An entity that is the smallest unit of interest in a complex dynamic. An agent could be a person, idea, task, knowledge, etc.

Aggregation Coevolutionary emergence of diverse ideas, agents, etc.

Aggregation mechanisms Macro-level complexity mechanisms. Refers to the dynamic interactions of perturbations (see initiating mechanisms), amplifications (expansion and elaboration of a perturbed state), and phase transition (nonlinear, form-shift in some part of an organization).

Coevolution A process in which "reciprocal selective pressures operate to make the evolution of ... (one agent in an interactive process) partially dependent on the evolution of (other agents)" [67]. Coevolution occurs when heterogeneous agents in a niche are interdependent such that they dynamically adjust to each others changes.

Commodity-based economy Refers to a manufacturing economy in which the major assets are physical commodities.

Complexity organizational knowledge Systemic knowledge that imbues organization with capacity for adaptive and creative responses; attributable to the strength and viability of the system's complex structures and to the viability of the relationship between the organization's complex structures and its bureaucratic structures.

Complexity mechanism Social mechanisms that underlie complexity dynamics.

Complexity theory In organizational sciences, the study of emergent dynamics in neural-like networks of adaptive, vision-oriented agents.

Conflicting constraints Conflict that emerges when the preferences of one agent challenges the preferences of another. In complex networks, conflicting constraints generate pressure to elaborate.

Dissipative structures Structures that emerge when far from equilibrium systems release excess energy in a phase transition.

Emergence The appearance of new structures or ideas from the actions of complex interactions.

Enabling behaviors Activities that foster conditions (e. g., interdependency, enabling rules, adaptive tension) in which complex dynamics can emerge.

Enabling rules Rules that govern interactions among adaptive agents in complex systems. Contrast with bureaucratic rules, which delimit the responsibilities of agents in a closed system bureaucracy.

Entanglement A dynamic relationship between the formal administrative forces and informal complexly adaptive, emergent forces of an organizational systems.

Equilibrium A stable, predictable relationship among agents and structures of a relationship. Related to thermodynamic concept of a low energy state.

Extreme event Instances of dramatic change that occur infrequently in social organizations; the US government's experience of the Katrina hurricane in New Orleans illustrates. Extreme events typically require rapid action and can be minimally responsive to top-down control.

Far from equilibrium Typically defined as a high energy state; defined here for organizational studies as an intensely complex, dynamic state driven by excess levels of pressure and perturbations.

Heterogeneity A diversity of skills, worldviews, preferences, beliefs, goals, (etc.) among interactive agents in a complex system.

Initiating mechanisms Micro-level complexity mechanisms. They include coevolutionary interaction and perturbations (an unexpected change in an interaction relationship; attributable to complex, interactive dynamic).

Interdependency Network conditions in which the actions of one agent are influenced by the actions of another.

Knowledge economy Refers to a production economy in which the major assets are the knowledge possessed by individuals and networks of individuals (organizational knowledge).

Meso theory Variously defined as theory that bridges macro and micro level theory; different levels of hierarchy; or different levels of analysis (individual, dyadic, group).

Multi-agent based modeling Agent based modeling procedure that analyzes different types of networks simultaneously (e. g., agent networks, task networks, etc.)

Organizational level From Jaques [40], the level of bureaucracy that lies between the upper echelon levels and the work production level; includes middle management.

Perturbations Events that disturb normal organizational interactions and generate pressures that can lead to phase transitions.

Phase transition Sudden, nonlinear form shifts attributable to the dissipation of accumulated pressure in complex systems. Phase transitions can occur at multiple levels of intensity.

Postmodernism In general, a rejection of scientific modernism. For organizational science, it represents a realization that organizational behaviors cannot be adequately expressed as mathematical relationships among variables. Complexity theory adds that organization is an ultimately unpredictable dynamic whose causal structure is based on interactions among complexity mechanisms.

Production level From Jaques [40], the level of bureaucracy responsible for line production.

Requisite complexity McKelvey and Boisot's [65] modification of Ashby's requisite variety; they maintain that viable organizations are at least as complex as their competition.

Requisite variety Ashby's [5] dictum that viable organizations have at least the same degree of flexibility as their competition.

Spaces between A proposal that creative ideas emerge from interactive dynamics. Creative emergence is generated when agents work to resolve tension; such tension is product of conflicting constraints, heterogeneous preferences, ideas, and knowledge, etc.

Social mechanism A process attributable to dynamic interactions among multiple people, variables, ideas, physical limitations, etc. Mechanisms describe "a set of interacting parts – an assembly of elements producing an effect not inherent in any one of them" p. 336 in [27], p. 74 in [37].

Strategic level From Jaques [40], the upper echelon level of bureaucracy; responsible for organizational strategy, policy making, resource acquisition and allocation, etc.

Swarm behavior Study of the dynamics of swarms in biology and in organizations. Scientists are finding that amazingly complex, "intelligent" behavior can emerge from complex behaviors that are structured by a few, simple enabling rules. Useful in organizations for improving such things as distribution efficiency or flexible response to environmental conditions.

System dynamics Network simulations in which relationships among agents are defined mathematically.

Top-down administration Administrative practices in which decisions are made by superiors to be carried out by subordinates.

Vision A teleological view of the future. Vision can be highly specific, or determinate, as when an organization projects future markets (such visions are more properly, goals; see Sharp Corp.'s strategy statement in this paper for an example). Vision statements appropriate for complex systems are indeterminate in that they do not preclude the future; for example, "This company will strive to enhance its competitive advantage by optimizing its flexibility."

Definition of the Subject

Complexity theory for organizations examines the influence of complex dynamics on (among many things) organizational structure, leadership, power and control, influence, and strategy. It is applicable not only to understanding but of practice in any organizational type whose primary commodity is knowledge and application of knowledge. The core outcomes of complex dynamics are creativity, adaptability, and learning. Complexity theory is particularly germane in what has come to be called, the knowledge economy.

This paper defines basic premises underlying complex dynamics in organizations, and argues in particular that organizational complexity is best understood as the interactions among complexity mechanisms (causal process attributable to dynamic interactions among multiple people, variables, ideas, etc.) rather than as the outcome of defined variables. We explore the relationship between bureaucracy and complexity, and define three levels of behavior in complex organizations: administrative, enabling, and adaptive. We apply these ideas along with complex premises underlying complexity to strategic leadership. The paper describes how complexity produces its core outcomes – creativity, adaptability, and learning – and argues that these outcomes are particularly pertinent to today's economy. It concludes with a discussion of unique research methodologies that can be used to study a process based on mechanisms rather than on variables.

Introduction

The science of complexity offers interesting solutions to some difficult challenges in organizations. These challenges affect nearly all economic sectors of society, from electronics to education, computer programming to automotives, banking, government, and the armed forces. They include hyper-competitive environments [38], globalization, unequal global distribution of wealth and production costs, rapid technological and social change, highly complex operations (e. g., the complexity of sophisticated software programming), and the changing nature of the workforce (less loyal, comfortable with electronic communication, demanding of flexible work arrangements). To address these challenges, practitioners and scholars recognize the need for radically new leadership paradigms [81], revised perspectives of organizational strategy [81], flexible structures [83], rapid adaptability [16], and decentralized problem-solving structures [25]. Complexity science is applicable to these challenges because, among other things, it focuses on flexible, interactive dynamics rather than top-down control, it offers unique descriptions of change in dynamic system, it reveals important clues about creativity, adaptability, and learning in organizations, and it shifts attention from central tendency variables to causality based on dynamic mechanisms [50,81].

Traditional thought in organization science was oriented toward efficient production and top-down coordination. It assumed that organizational behavior was intimately dependent upon structure provided by positional leaders, rational plans and goals, centralized coordination, organizational evolution, and vision [35]. It has explored the nature and impact of power [41], institutional forces [28], and charisma [23]. Traditional thought has, like complexity theory, described change but has assumed that organizational change and innovation must be planned, that effective action must be coordinated, and that organization must be structured in the board room.

By the 1990s, this "top-down" notion began losing credibility among organizational scholars and practitioners (e. g., [35]). The major culprit in this demise was the knowledge economy – the globalization, rapid change, and hyper-competitive environments described above. Top-down perspectives of organization had been oriented toward more stable, commodity based environments [9], and the emerging knowledge economy is oriented toward creativity, adaptability, and learning.

Complexity theorists in the organizational sciences seek relevant strategies for dealing with the knowledge economy. Complexity theory is useful for this because it does not merely improve traditional perspectives of doing organization but rather addresses different sets of behaviors than are typically addressed by organizational theorists. In particular, it focuses on informal interactive processes within an organization, on processes that emerge from informally interacting people and groups, and on productive relationships between informal and formal activities within an organization. It asks such questions as, "How can organizational decision making and interactive patterns be restructured and reorganized to increase flexibility" [49]? "How can leadership be re-conceptualized in a way that is effective for complex organizations" [70]? Or, "How do we re-focus organizational strategy for a knowledge economy" [58]?

In this paper, we look first at basic premises underlying the application of complexity theory to organizations, then look at how complexity can inform understanding of bureaucracy and strategic leadership. We examine the organizational outcomes of complex dynamics (hence the rational for applying this science), and conclude with a discussion of research methodologies that are useful for studying complex organizational behaviors.

Basic Premises

Studies have found that, like biological [30], economic [4], and physical [45] systems, emergent dynamics are readily observed among interacting people, or agents [20,68,69]. Interactive dynamics and emergence are common denominators in complexity studies, and while there are discipline-specific variations (organizational agents, for example, make learned, adaptive choices while weather systems react primarily to the interactions of physical properties), all are shaped by these defining dynamics. Agents interact and the interaction changes them (emergence).

Complexity organization researchers are interested in understanding how these dynamics operate in human systems. Lichtenstein and Plowman [51] have found, for example, that certain interactions among individuals and groups lead to perturbations, or pressure to elaborate. Consistent with Kauffman [43], these tensions pressure the system to adapt and to change. Perturbed, interactive systems also tend toward self-organization [57,62]. This phenomenon was first systematically observed in the 1940s and 1950s as informal groups [39] and was expanded by complexity theorists to include any situation in which ideas or people spontaneously, and without external motivation, act in some degree of synchrony with one another. Chiles et al. [20] identified dissipative structures – fluctuation, positive feedback, stabilization and recombination – in the emergence of the music industry at Branson, Missouri.

Researchers are also interested in understanding the conditions that enable emergent, self-organizing behavior in social systems [82]. *Interaction*, of course, is *sine qua non*. Schreiber, Marion, Uhl-Bien, and Carley [74], drawing from Kauffman [43], found that interactive agents must possess a requisite level of *interdependency* to enable learning outcomes (too much or too little will suppress complex dynamics; [43]). This interdependency generates *conflicting constraints*, which are important sources of perturbations in a complex system. McKelvey [64] adds that such pressure is also generated by external sources such as leadership behavior or economic pressures; he refers to this as *adaptive tension*. Uhl-Bien et al. [82] adopted this term to refer to both internal and external pressures.

Bonabeau and Meyer [10] have developed fascinating insights into the nature of swarm behavior in human organizations, and have used it to solve efficiency problems involving such things as package delivery and local surges in phone traffic [10]. Swarms are a function of a set of simple, *enabling rules*, such as "fill orders [in a distribution warehouse] until the next person in line takes over", or "drop 'digital' pheromone trails for subsequent messages to track" (in computerized instructions for phone systems).

Complex behaviors in social systems are enabled by *heterogeneity* of skills, preferences, outlooks, etc. Traditionally, organizational theorists, particularly those who study leadership, have sought to align employees to centralized goals [56]; complexity theory suggests that organizations whose employees pursue a diversity of goals (within the context of an interactive, interdependent network) will have a diversity of ideas to draw from as they seek creativity and adaptable changes.

Finally, Schreiber, Marion, Uhl-Bien, and Carley [74] found in a multi-agent based simulation that moderate levels of *vision* (or general organizational focus) are conducive to complex behaviors. High centralization of vision (in combination with high interdependency among agents) was counter-productive for complexity outcomes in their simulation, as was low levels of vision (low vision was more problematic to outcomes than was high vision).

Emergence and Phase Transitions

Emergence refers to the generation of new structures or ideas due to the actions of complex interactions. This phenomenon is important because it allows the system to rapidly produce adaptive, creative responses to environmental conditions. It occurs in social structures when heterogeneous agents or ideas act in relative synchrony with one another. The emergence of riot conditions out of the behaviors of just a few people illustrates (the emergence of a riot also illustrates that emergence is not necessarily a functional response to environmental contingencies; the notion of counter-productive emergence has been discussed by Uhl-Bien, Marion and McKelvey [82]).

One form of emergence involves the *accumulation* of different ideas or agents. Anderson [1] and Marion [55], for example, described the emergence of technology that led to the invention, in 1975, of the microcomputer. It began with disparate advances – transistors, printed boards, LEDs, etc. Engineers began using these isolated components together, as they did in the transistor radio popular in the 1960s. Accumulation took a big leap with the introduction of the handheld calculator in the late 1960s, which added such things as processors and information storage capacity. The microcomputer was the ultimate "accumulation" or emergence event.

Importantly, the components of such emergence processes accumulate because they "coevolve." Coevolution is defined by Pianka, p. 329 in [67], as a process in which "reciprocal selective pressures operate to make the evolution of either … [agent in the interactive process] partially dependent on the evolution of the other" see also McKelvey [64]. Coevolution occurs when heterogeneous agents in a niche are interdependent such that they dynamically adjust to each other's changes. It raises the important observation that an organizational species' fitness – the capacity to survive and thrive – is influenced by supporting roles from other organizational species [43].

A second form of emergence is called *phase transition*. Phase transitions are sudden, nonlinear form-shifts (changes in structure, knowledge, worldview, etc.). The sudden demise, in 1989, of the USSR illustrates this phenomenon. Phase transitions like that experienced by the USSR result from a build up of tension or pressure [32,64,71]. Haken [32], for example, illustrated phase transition in terms of slowly heated oil; eventually the heat (tension) reaches a point at which the oil precipitously transitions to a new state, observed as a gentle boil.

Phase transitions, occur at multiple levels of intensity, and a number of researchers have observed that, in complex systems, intensity of transitions plotted against their frequency describes a power law relationship ($1/f^x$); [2, 7,75]. Power law curves are steeply descending slopes that quickly bottom out into a more gently descending slope; when intensity and frequency are converted to log equivalents, the slope becomes a straight line descending left to right like a negative regression line in statistics. Power law curves describe situations in which high intensity events occur infrequently while low intensity events occur often.

This has opened a line of exploration into what are called, extreme events, or analyzes of phenomena that are highly intense but of low frequency (the 2008 Organization Science Winter Conference was devoted to this topic). Stephenson and Bonabeau [76] concluded, for example, that bottom-up, or adaptive, leadership strategies may be more effective in extreme events such as Katrina or the Twin Tower attacks of 9/11 than are top-down leadership strategies.

Far from Equilibrium

The accrued tension described by Haken generates what Prigigine [71] calls, a far from equilibrium (FFE) state. At FFE; the system will often act spontaneously to release (or dissipate) built up tension by moving to a new structural state [64].

In organizational studies, the far from equilibrium notion is unique, for organizations have traditionally been described in terms of equilibrium structures ([12]; see Maruyama [60], for a notable exception). The idea that far from equilibrium states exist suggests an equilibrium continuum for organizations that stretches from an equilibrium state to far from equilibrium conditions. Hazy and Marion [34] argue, however, that while this may be true for some physical systems, it is not true of human systems. Human systems always function in a state of change, albeit with varying degrees of complexity, and it makes no sense to juxtapose such systems against an equilibrium end point. They conclude, then, that far from equilibrium in a social system occurs when the dynamical conditions and resulting complexity of a social system are increased beyond a certain point, not when it moves away from equilibrium.

Even so, the assumption that organization exists in a state of equilibrium is common in the traditional literature on organizations. Although it has been clarified by earlier organizational theorists to be a moving equilibrium (called homeostasis) that adjusts to environmental conditions (the influential classics of this argument are [12]; and [84]), the basic premises in organizational studies have implicitly or explicitly equated organization with adapted variations of the equilibrium view from thermodynamics. Complexity theory contradicts this notion, arguing instead that adaptive systems can only survive and thrive in a state variously called far from equilibrium [64], complexity, edge of chaos [42], or Type IV order [47]. If Hazy and Marion [34] are correct, perhaps the better terms for this are those that do not imply an equilibrium state at all.

Complexity Mechanisms

Complexity is unique in organizational studies because it is oriented around social processes rather than around variables. A variable, or measure of central tendency, is static and it reveals little about the evolutionary social processes in which the variable is embedded. Complexity seeks to understand and to describe the behaviors of those social processes, or what Davis and Marquis [27] and Hedström and Swedberg [36] and others have labeled, social mechanisms.

Mechanisms are defined as commonly observed patterns of behavior [29]; they describe "a set of interacting parts – an assembly of elements producing an effect not inherent in any one of them" p. 336 in [27]; see also p. 74 in [37]. Mechanisms are processes, or dynamic interactions among agents, ideas, and structures. Elster's [29] definition proposes that certain mechanisms can be "commonly" observed across multiple phenomena (that is, they are not necessarily unique to given conditions). They illustrate by arguing that the logistic curve (emergence begins slowly, picks up speed, then levels off) is a mechanism and that it can be observed in numerous situations (Mozart's lifetime productivity, emergence of product popularity, etc.)

Uhl-Bien and Marion [81] identify two broad, interdependent categories of universal mechanisms that they call, complexity mechanisms: These are initiating mechanisms, or interactive processes that generate perturbations, or disturbances, in a system, and aggregation mechanisms, which include accumulation mechanisms (accruing of ideas or structures) and phase transitions (described earlier as nonlinear changes that generate new order). Initiating mechanisms are more micro level while aggregation mechanisms are more macro level.

Practitioners who grasp this notion of complexity mechanisms do not perceive organization in terms of average motivation or efficiency, but rather perceive organization as sets of interacting, evolving processes. They do not ask, for example, whether satisfaction will increase if leaders engage in certain specific acts, or whether high leadership scores on a transformational scale will enhance productivity. Rather they want to know how interactive, interdependent mechanisms influence the capacity of social networks within organizations to generate unpredictable but creative outcomes. These contrasting sets of illustrative questions differ in that the former are based on scientific assumptions about linear relationships among central tendencies and the latter are based on postmodern assumptions regarding complex interactions that are indeterminate in the detail but determinate in the general nature of their outcome. For example, we can predict dynamic creativity capacity but we cannot predict what that creativity will look like.

Understanding Organization with Complexity

We turn now to an examination of some specific applications of complexity to organizational theory and behavior. This discussion is framed around the relationship of complexity to bureaucracy and around strategic leadership.

Bureaucracy and Complexity

The discussion to this point has implicitly portrayed organization as an adaptive dynamic, or as a system driven exclusively by agents acting informally within complex networks to produce emergent outcomes. In reality, organizations are bureaucratic structures, and despite predictions by some of a post-bureaucratic economy [31,35,54], it is unrealistic to ignore this pervasive element of organization (as Weber [85], said, bureaucracy will be with us until the last shovel of coal is dug from the earth).

For this reason, complexity theorists Uhl-Bien et al. [82] have chosen to depict informal, complex dynamics as embedded within each of Jaques' (1989) three levels of bureaucracy: production, organizational, and strategic. They observe that, while bureaucratic functions differ at each level, the nature of complex dynamics differ only in the tasks they address. That is, complex behavior is similar across the system, whether observed in the board room or in the production divisions. Dynamic, complex behaviors are pervasive, and are universally driven by the enabling conditions described in the last section. Specific complex processes are entangled with one another and even across bureaucratic level. Further, the inter-influences across these dynamics is more horizontal than hierarchical: Dynamics on the shop floor – Jaques' [40] production level – can directly and quickly influence dynamics in the executive suite – Jaques' strategic level. Using complexity terminology, complex dynamics in organizations are, therefore, fractal in nature. From an organizational perspective, they represent a meso view of organization; that is, they describe the linkages that unite micro, or individual, level structures with macro, or organizational level, analysis [26]. Uhl-Bien and Marion [81] also argue that they are meso in that they link different hierarchical levels.

From this, Uhl-Bien, Marion, and McKelvey [82] propose three key dynamics within organizations (which they describe relative to leadership): administrative, adaptive, and enabling. *Administrative leadership* is related to traditional perspectives of leaders as those who generate organizational strategy, gather and distribute resources, and create policy. They differ from traditional perspectives in that their actions must serve to nurture the somewhat fragile, bottom-up (or adaptive) complex leadership processes (fragile in that they are somewhat sensitive to political power behaviors).

Adaptive leadership is defined as any interactive behavior that creates a perturbation which leads to a phase transition. The key elements are interaction, perturbation, and phase transition. Uhl-Bien et al. argue that change originates in the "spaces between" interacting agents (see also [52]); that is, it is the product of the push and shove that can occur between or among agents with heterogeneous skills or worldviews. Perturbations are resulting disturbances in a network (change, new ideas or perspectives, inventions, etc.) that reshape the interaction patterns in that network [81]. Phase transitions are nonlinear shifts within a network (these are described more fully below).

Enabling leadership, the third of Uhl-Bien et al. [82] types, refers to individuals who foster the enabling conditions described above (interaction, interdependency, heterogeneity, enabling rules, etc.) and who mediate the relationship between adaptive and administrative functions.

Emergence, Niche, and Strategic Leadership

The earlier discussion of emergence and phase transitions help us re-visualize the role of another strand in organization sciences: strategic leadership. Strategic leadership theorists study upper echelon behaviors of leader/agents responsible for positioning an organization in its environment. Traditional literature has perceived strategic leadership as a top-down, goal oriented process [13] in which the leader manages the organization's environment. For example, Sir Howard Stringer articulated the strategic goal of Sony Corporation at the 2006 International Consumer Electronics Show in Las Vegas, as:

> Sharpening our focus, shattering internal silos, streamlining our product offerings and growing ever more consumer-centric. We have analyzed our product lines and reorganized our corporate structure to become more nimble and better able to deliver champion products and a focused product line-up. [78].

The website for Sharp Corporation articulates both its philosophy and its business strategy in terms of expanding its market (sharp-world.com/corporate/info/philosophy/index.html). Strategic goals from other companies could be cited, but most will, like these, focus on the company's competitive position in its environment.

Given such assumptions, traditional strategic leadership literature has examined such things as the personality and decisions of top echelon leaders, relationships between the CEO and his or her board of directors, transformational leadership, and visionary leadership (see [13], for a summary).

More recent work by strategic leadership scholars is still concerned about position in the environment, but has focused not on upper echelon behavior but on the internal dynamics of organizations and how organizational flexibility contributes to organizational strategy [8]. Interestingly, Sharp has heeded this call: Stringer's articulation of their goal (above) includes a call for a "more nimble" firm. However, a review of the complexity literature has revealed little that has addressed this issue, despite the fact that it seems a natural subject for complexity analysis. One exception is an article by Marion and Uhl-Bien [58], which argues that flexible, complex structuring is key to organizational strategy. They propose that firms work to increase interaction, interdependency, and adaptive tension.

Marion and Uhl-Bien [58] also suggest that that traditional external-oriented strategic leadership has missed an important point. Much of this literature is, at least implicitly, about dominating the environment or beating the competition. Complexity theory argues that different organizations coevolve [43], and in doing so, they grow to depend upon one another. That is, coevolving systems aggregate and form a niche, or interdependent infrastructure upon which each part of the niche depends to some extent. Automobile makers depend upon fuel outlets and automotive repair shops, for instance – and vice versa. Automotive makers even depend on one another, for their competitive relationship generate pressures that foster structural and ideational elaboration and generate a heterogeneous set of creative solutions that foster even greater creativity. External strategy, then, may be about more than competition; it may also be about fostering effective infrastructure: Marion and Uhl-Bien called this, survival of the cooperative.

Complexity Outcomes: Creativity, Adaptability, Learning

From all this, we can generalize the outcomes of complex dynamics for organizations as creativity, adaptability, and learning. *Creativity* is defined as an ideational phase transition in which an outcome reasonably could not have been predicted based on preexisting conditions or paradigmatic assumptions. Described more basely, it is something that "no one saw coming." Since it is a phase transition, it can be a minor bit of creativity or a major insight.

The major creative insights, of course, capture our imagination and dramatically change our way of seeing things (e. g., Einstein's relativity or Schrödinger's cat). However, small creative events are more common and, arguably, just as important. Creative ideas exist within complex networks that enable them to coevolve and aggregate. That is, they become part of the emergence dynamic within an organization. Creative ideas and associated structures change, aggregate with other creative ideas or structures, and generate adaptive pressures that maintain or increase the complexity of a system. Without the pressures that creative ideas generate, an organization would move towards increased isomorphism and decreased complexity – they move to a state that some would mistakenly call, homeostasis.

Adaptability is enhanced when an organization increases its level of complexity. This translates into (among other things) increased heterogeneity of skills and preferences. Traditional organizational theory has assumed that everyone should be on the "same page" (e. g., "Leaders … are responsible for the dissemination of strategic organizational goals, as well as for convincing their constituents to effectively implement those goals" p. 626 in [6]). Complexity theorists argue that, although organizational agent should be united by a vision, dynamic heterogeneity enables organizations and agents to rapidly and effectively deal with environmental exigencies. Ashby [5] refers to this as "requisite variety;" he defines it by stating that it takes variety to defeat variety. McKelvey and Boisot [65] call it "requisite complexity," thus underscoring the dynamic, co-evolving nature of complex systems and their adaptive response capabilities.

Learning is defined for organizations in several ways. Many of us would define it as something that an individual does. Some organizational theorists [3,22,66], including some complexity scholars (e. g., [9]), argue that learning is best understood as an organizational process. The functioning of the brain exemplifies this logic (see [44]). Individual neurons "learn" only small pieces of an environmental event; it is only when those pieces are put together within a network of pieces that we began to make sense of that event. This, crudely, is what is meant by organizational learning. Organizational knowledge can be defined as a network of different but synchronized bits of individual knowledge. Organizational learning is related to the complex nature of those networks – that is, to its dynamic, changing capabilities. If, then, a system increases its complexity, it will increase its capacity to learn.

Still others propose that organizational learning is a function of the capacity of the network to disseminate information [15]. Like the complexity perspective of organizational knowledge in the previous paragraph, this perspective is intimately linked to the degree of complexity in the organizational network.

Methodology for Studying Complex Organizations

Before developing the methodological strategies that are useful for studying complex organizational system, it will be helpful to summarize a few of the assumptions that underlie complexity, all of which have been stated or implied in this manuscript. These assumptions guide our selection of methodological strategies in organizational studies.

1. Complexity science examines the dynamics of interactions among large numbers of adaptive agents [24].
2. Complex systems evolve dynamically and change nonlinearly (unpredictably; [21]).
3. Human systems do not exist in a state of equilibrium.
4. Complex systems are best described in terms of interactions among complexity mechanisms rather than as interaction among variables.
5. Complex dynamics are enabled rather than planned or controlled.

Traditional Methodologies in the Social Science

Perhaps the most common methodology used to study organizations involves field research strategies and regression statistical procedures. Field strategies permit researchers to draw conclusions in non-laboratory conditions. Regression techniques are based on assumptions of regular, predictable relationships among variables; if one finds such a relationship between variable A and variable B, the relationship was true yesterday and it will be true tomorrow [61]. Such relationships are predictable in that they assume that if you increase variable A, then a resulting change in variable B can be calculated.

While statistical methodology is certainly useful, it cannot deal with many of the conditions posed by complexity science. It is not suited, for example, for exploring dynamic aggregation in organizations, nor is it helpful for explaining nonlinear changes that occur at far from equilibrium states. In particular, its usefulness lapses when one seeks to understand interactions among, and outputs of, complexity mechanisms. Complexity mechanisms are a function of dynamic interactions among numerous agents, ideas, and even variables. Traditional statistical methodology is just not suited for this type of analysis. Statistical assumptions are better suited to equilibrium systems in which desired outcomes involve planning and control. Complexity science is not particularly interested in such outcomes.

A growing trend in organizational studies involves the use of qualitative strategies. Qualitative research requires careful observation of a social dynamic followed by thorough analysis of what was observed. Qualitative methodology is useful for observing dynamic activities and social mechanisms, and is rather widely used for that purpose [20,49,68,70]. Its drawbacks are that it is time and resource intensive and it does not allow one to freely experiment with, or freely change, a social mechanisms – one cannot readily play what – if games in qualitative analysis.

Methodologies for Complexity

The most widely used methodologies in complexity study in organizations are agent-based modeling (ABM), systems dynamics, and qualitative research. ABM replicates or mimics interactive dynamics among organizational agents. ABM is a simulation procedure in which individually programmed agents interact in a computer according to certain rules. Commonly used examples are NetLogo by Uri Wilinsky, Dynamic Network Analysis by Kathleen Carley, and Repast Agent Simulation Toolkit by Michael J. North. These simulations allow researchers to simulate interactions under a variety of conditions and to test the effects of various changes in those conditions. ABM has been used to explore such things as the dynamics that underlie personnel turnover [14], organizational design and performance under stress [19], and leadership as enabler of complex organizational functioning [73].

System dynamic analysis is similar to ABM in that it examines evolving relationships among agents. It differs in that those relationships are not governed by social rules as they are with ABM's, rather they're governed by mathematical defined relationships. Like ABM, it permits the researcher to experiment with various conditions. Sterman [77], for example, has used it to examine interactions among individual decisions and the effects of feedback on the decision process.

Both ABM and system dynamic computer programs permit researchers to base their simulations on data collected in real-world situations. Once that data is entered, however, the quality of the subsequent simulation is only as good as the quality of the rules or the equations that govern it. However the intent is not to predict specific outcomes; rather, the intent is to understand how dynamic relationships can affect outcomes. Carley's Dynamic Network Analysis, for example, permits a user to evaluate the effects of an evolving network on the capacity of the organization to diffuse knowledge. One can play with such things as the degree of interdependency among agents or the degree of centrality around vision (as Schreiber et al. did in their [74] study) to see how they influence the learning dynamic.

Qualitative research performs generally the same things that ABM and system dynamic analysis do, but it allows the researcher to continuously examine real life events rather than merely simulating those dynamics. The benefit is a closer description of reality; the drawbacks are that it is highly resource and time intensive (especially if one wants to obtain results as in-depth as are results available through simulations), and it does not generally permit much control over conditions that might influence outcomes. Interesting qualitive studies of complex dynamics have been conducted by such organizational theorists as Plowman, Baker et al. [68], who examined emergent conditions in an inner-city church, and Chiles, Meyer, and Hench [20], who studied the complex evolution of Branson, Missouri.

Conclusions

I believe that the best definition of complexity behavior is provided by Paul Cilliers [21]. To frame that definition, he argues that we often confuse complicated with complexity. A complicated system is one with many components interacting to perform complicated tasks, but all of those components maintain their original integrity. A jet plane, he said, is complicated: the components of the plane are unchanged by their relationship. Complex systems are composed of many interacting units, all of which are changed by that interaction such that they quickly loose their original form and function. Organizational theorists who study complex dynamics are seeking to observe and explain that dynamic [48,63]; to understand how organizations might be restructured to improve their capacity for complex outcomes [11]; to improve general leadership in organizations [51,57,80]; to restructure our understanding and implementation of strategic leadership [58], and to restructure how we think about reality, understanding, predictability, and scientific methodology [16,17,18,53,72].

Complexity is genuinely an interdisciplinary science, applicable equally to biology, physics, and the social sciences. Organizational sciences have experienced a number of cumbersome imports in the past: systems theory borrowed from biology [12], as did population ecology theory [33], and scientific management theory borrowed from mechanics [79]. Complexity theory, however, is not about biology or mechanics, it is about interaction dynamics – and this is common across multiple phenomena. It is also about change, not gradual change but change that surprises. It helps us understand how systems learn and adapt to their environment. It comes at a good time for organizational studies, and will help us adapt organizations to the dramatic, knowledge-oriented changes being experienced in world economies.

Future Directions

One is tempted to argue that organizational complexity theory represents a paradigm shift in the way we think about reality (in the Kuhnian sense; [46]), and there is considerable reason to do so. Complexity theorists, for example, acknowledges that the problems facing organizations are far too complex to be solved by rational thinking and planning, and that organizations must engage the interactive, complex capacity of all its employees to solve those problems [59]. Because of this, organizational complexity theory is probably one of the first truly postmodern theories of organization. Organizational complexity theory is premised on complexity mechanisms rather than on variables, thus the questions that emerge from this perspective are dramatically different from any that have been asked in the past (e. g., "how does level of interdependency interact with level of vision to influence temporal emergence of, and phase transitions in, organizational learning"). Complexity theory clearly separates itself from the notions of equilibrium that have so influenced 20th century organizational theory. It moves us away from the centrality of administration and focuses on the centrality of bottom-up, adaptive leadership. Its notions of leadership do not revolve around interpersonal relationships, as does most of leadership theory in the 20th century; rather, it revolves around interactive dynamics. Complexity theory proposes that organizational strategy should seek effective relationships, rather than competition, with other members of its niche. Simply put, complexity theory is a dramatically different way to understand organization.

Complexity theory has captured the attention of academics and practitioners. Important indications of this include the large number of articles on the subject that are appearing in scholarly journals (e. g., the August, 2007, edition of *The Leadership Quarterly* was a special issue on complexity leadership) and the fact that complexity-related articles are being published in top-tier journals (e. g., [11,68]); the increasing number of complexity-related grants in organizations (NSF began a complexity category for organizational scholars in 2007); and a significant surge in the number of practitioners who are seeking consultant relationships with complexity theorists.

A number of interesting research questions are emerging from the complexity paradigm. Andriani and McKelvey's work on extreme events could have significant implications for the way we approach Katrina-like tragedies in the future, for example. The work to develop methodologies such as agent based modeling can have tremendous impact on how we analyze organizational dynamics. The idea that organizational behavior is about the actions of mechanisms is beginning to sink in among scholars, and could dramatically realign the way we think about organization. Complexity leadership theory poses unique questions for the role of strategic leadership and challenges the very way we perceive leadership.

Arguably, there has not been a true paradigm shift in organization studies since Weber introduced his concept of bureaucracy [59]. At the risk of being overly optimistic, organizational complexity theory may be changing this.

Bibliography

1. Anderson P (1995) Microcomputer manufacturers. In: Carroll GR, Hannan MT (eds) Organizations in Industry. Oxford University Press, New York, pp 37–58
2. Andriani P, McKelvey B (2006) Extremes and scale-free dynamics in organization science: Some theory, research, statistics, and power law implications. Manuscript submitted for publication
3. Argyris C, Schön DA (1978) Organizational learning: A theory of action perspective. Addison-Wesley, Reading
4. Arthur WB (1989) The economy and complexity. In: Stein DL (ed) Lectures in the sciences of complexity, vol: 1. Addison-Wesley, Redwood City, pp 713–740
5. Ashby WR (1960) Design for a Brain, 2nd edn. Wiley, New York
6. Berson Y, Avolio BJ (2004) Transformational leadership and the dissemination of organizational goals: A case study of a telecommunication firm. Leadersh Q 15(5):625–646
7. Blau JR (1995) Art museums. In: Carroll GR, Hannan MT (eds) Organization in Industry. Oxford University Press, New York, pp 87–114
8. Boal K (2004) Strategic Leadership, Organizational Learning and Network Ties. Paper presented at the Conference on Strategic Leadership on Both Sides of the Atlantic, International Institute for Management Development, Lausanne, 2004
9. Boisot M (1998) Knowledge assets: Securing competitive advantage in the information economy. Oxford University Press, Oxford
10. Bonabeau E, Meyer C (2001) Swarm intelligence: A whole new way to think about business. Harvard Bus Rev 79(5):107–114
11. Brown SL, Eisenhardt KM (1997) The art of continuous change: Linking complexity theory and time-paced evolution in relentlessly shifting organizations. Adm Sci Q 42(1):1–34
12. Buckley W (1967) Sociology and modern systems theory. Prentice Hall, Englewood Cliffs
13. Canella AA, Monroe MJ (1997) Contrasting perspectives on strategic leaders: Toward a more realistic view of top managers. J Manag 23(3):213–230
14. Carley K (1992) Organizational learning and personnel turnover. Organ Sci 3:20–46
15. Carley K (1996) Organizational learning and personnel turnover. In: Cohen MD, Sproull LS (eds) Organizational Learning. Sage, Thousand Oaks
16. Carley K (1997) Organizational adaptation. An Oper Res 75: 25–47
17. Carley K (1999) On the evolution of social and organizational networks. In: Bacharach S, Knoke D, Andrews SB (eds) Research in the Sociology of Organizations: Networks In and Around Organizations, vol: 16. Press JAI, Greenwich, pp 3–30
18. Carley K (2002) Intra-organizational complexity and computation. In: Baum JAC (ed) The Blackwell Companion to Organizations. Blackwell, Oxford, pp 208–232
19. Carley K, Lin Z (1995) Organizational designs suited to high performance under stress. Trans IEEE Syst Man Cybern 25:221–230
20. Chiles T, Meyer A, Hench T (2004) Organizational emergence: The origin and transformation of Branson, Missouri's musical theaters. Organ Sci 15(5):499–519
21. Cilliers P (1998) Complexity and postmodernism: Understanding complex systems. Routledge, London
22. Cohen MD, Sproull LS (1996) Organizational Learning. Sage, Thousand Oaks
23. Conger JA (1999) Charismatic and transformational leadership in organizations: An insider's perspective on these developing streams of research. Leadersh Q 10(2):145–179
24. Coveney P (2003) Self-organization and complexity: A new age for theory, computation and experiment. Paper presented at the Nobel symposium on self-organization, Karolinska Institutet, Stockholm, 2003
25. Cusumano M (2001) Focusing creativity: Microsoft's "Synch and Stabilize" approach to software product development. In: Nonaka I, Nishiguchi T (eds) Knowledge emergence: Social, technical, and evolutionary dimensions of knowledge creation. Oxford University Press, Oxford, pp 111–123
26. Dansereau F, Yammarino FJ (eds) (1998) Leadership: Multiple-level approaches: Contemporary and alternative, vol: 2. Press JAI, Stamford
27. Davis GF, Marquis C (2005) Prospects for organization theory in the early twenty-first century: institutional fields and mechanisms. Organ Sci 16(4):332–343
28. DiMaggio PJ, Powell WW (1991) The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. In: Powell WW, DiMaggio PJ (eds) The new institutionalism in organizational analysis. University of Chicago Press, Chicago, pp 63–82
29. Elster J (1998) A plea for mechanisms. In: Hedström P, Swedberg R (eds) Social Mechanisms. Cambridge University Press, New York, pp 45–73
30. Goodwin B (1994) How the leopard changed its spots: The evolution of complexity. Charles Scribner's Sons, New York
31. Grey C, Garsten C (2001) Trust, Control and the Post-Bureaucracy. Organization Studies, vol: 22. Walter de Gruyter, New York, pp 229
32. Haken H (1983) Synergetics, an introduction, vol: 3. Springer-Verlag, Berlin
33. Hannan MT, Freeman J (1977) The population ecology of organizations. Am J Sociol 82:929–964
34. Hazy J, Marion R (2008) When Organizing in Extreme Situations, Is Far-From-Equilibrium (FFE) a Meaningful Concept?

Poster presented at the Organization Science Winter Conference, Squaw Creek, 2008

35. Heckscher C, Donnellon A (1994) Defining the post-bureaucratic type. New perspectives on organizational change Thousand Oaks: Sage, In: Heckscher C, Donnellon A (eds) The post-bureaucratic organization
36. Hedström P, Swedberg R (1998) Social Mechanisms: An analytical approach to social theory. Cambridge University Press, Cambridge
37. Hernes G (1998) Real virtuality. In. Hedström P, Swedberg R (eds) Social mechanisms: An analytical approach to social theory. Cambridge University Press, Cambridge, pp 74–101
38. Hitt MA (1998) Presidential Address: Twenty-first century organizations: Business firms, business schools, and the academy. Acad Manag Rev 23:218–224
39. Homans GC (1950) The human group. Harcourt, New York
40. Jaques E (1989) Requisite organization. Cason Hall, Arlington
41. Jermier JM (1998) Introduction: Critical perspectives on organizational control. Adm Sci Q 43(2):235–256
42. Kauffman SA (1993) The origins of order. Oxford University Press, New York
43. Kauffman SA (1995) At home in the universe: The search for the laws of self-organization and complexity. Oxford University Press, New York
44. Kelso JA (1995) Dynamic patterns: The self-organization of brain and behavior. Press MIT, Cambridge
45. Koschmieder EL, Davis S, Tvergaard V, Batchelor GK, Leibovich S (eds) (1993) Bernard cells and Taylor vortices–Cambridge monographs on mechanics and applied mathematics. Cambridge University Press, New York
46. Kuhn TS (1970) The structure of scientific revolutions, 2nd edn. The University of Chicago Press, Chicago
47. Langston CG (1986) Studying artificial life with cellular automata. Phys 22D:120–149
48. Lichtenstein B (2000) Dynamics of rapid growth and change: A complexity theory of entrepreneurial transitions. In: Liebcap G (ed) Advances in the study of entrepreneurship, innovation, and economic growth, vol: 6. Press JAI, Westport CT, pp 161–192
49. Lichtenstein B, Carter N, Dooley K, Gartner W (2007) Complexity dynamics of nascent entrepreneurship. J Bus Ventur 22:236–261
50. Lichtenstein B, Plowman D (2007a) Emergent leadership: A meso-model of leadership inspired by complexity. Manuscript submitted for publication
51. Lichtenstein B, Plowman D (2007b) The leadership of emergence: Meso-level leadership. Paper presented at the Festschrift for Jerry Hunt, Texas Tech, 2007
52. Lichtenstein B, Uhl-Bien M, Marion R, Seers A, Orton D, Schreiber C (2006) Leadership in Emergent Events: Exploring the Interactive Process of Leading in Complex Situations. Emerg Complex Organ 8(4):2–12
53. Lin F, Pai Y (2000) Using Multi-Agent Simulation and Learning to Design New Business Processes. Trans IEEE Syst Man Cybern Part A Syst Hum 30(3):380–384
54. Maravelias C (2003) Post-bureaucracy–control through professional freedom. J Organ Chang Manag 16(5):547
55. Marion R (1999) The edge of organization: Chaos and complexity theories of formal social organizations. Sage, Newbury Park
56. Marion R (2006) Complexity in organizations: A paradigm shift. In: Sengupta A (ed) Chaos, Nonlinearity, Complexity: The Dynamical Paradigm of Nature, vol: 206. Springer-Verlag, Berlin, pp 248–270
57. Marion R, Uhl-Bien M (2001) Leadership in complex organizations. Leadersh Q 12:389–418
58. Marion R, Uhl-Bien M (2007a) Complexity and strategic leadership. In: Hooijberg R, Hunt J, Antonakis J, Boal K, Lane N (eds) Being There Even When You Are Not: Leading Through Structures, Systems, and Processes, vol: 4. Elsevier, Amsterdam, pp 273–287
59. Marion R, Uhl-Bien M (2007b) Paradigmatic Influence and Leadership: The Perspectives of Complexity Theory and Bureaucracy Theory. In: Hazy JK, Goldstein J, Lichtenstein B (eds) Complex Systems Leadership Theory. Publishing ISCE, New York, pp 143–159
60. Maruyama M (1963) The second cybernetics: Deviation amplifying mutual causal processes. Am Sci 51:164–179
61. Maxwell JA (2004) Causal explanation, qualitative research, and scientific inquiry in education. Educ Res 33(2):3–11
62. McKelvey B (1999) Complexity theory in organization science: Seizing the promise or becoming a fad? Emerg 1(1):5–32
63. McKelvey B (2001) Energizing order-creating networks of distributed intelligence. Int J Innov Manag 5:181–212
64. McKelvey B (2008) Emergent Strategy via Complexity Leadership: Using Complexity Science and Adaptive Tension to Build Distributed Intelligence. In: Uhl M-Bien, Marion R (eds) Complexity leadership. Conceptual foundations, vol: 1. Information Age Publishing, Charlotte, pp. 225–268
65. McKelvey B, Boisot MH (2003) Transcendental organizational foresight in nonlinear contexts. Paper presented at the Conference INSEAD on Expanding Perspectives on Strategy Processes, Fontainebleau, France, 2003
66. Nonaka I, Takeuchi H (eds) (1995) The knowledge-creating company. Oxford University Press, Oxford
67. Pianka ER (1994) Evolutionary Ecology, 5th edn. HarperCollins, New York
68. Plowman D, Baker LT, Beck T, Kulkarni M, Solansky S, Travis D (2007) Radical change accidentally: The emergence and amplification of small change. Acad Manag J 50(3):515–543
69. Plowman D, Duchon D (2008) Dispelling the myths about leadership: From cybernetics to emergence. In: Uhl M-Bien, Marion R (eds) Complexity leadership. Conceptual foundations, vol: 1. Information Age Publishing, Charlotte, pp. 129–154
70. Plowman D, Solansky S, Beck T, Baker L, Kulkarni M, Travis D (2007) The role of leadership in emergent, self-organization. Leadersh Q 18:341–356
71. Prigogine I (1997) The end of certainty. The Free Press, New York
72. Schreiber C, Marion R, Uhl-Bien M, Carley K (2006) Multi-Agent Based Simulation of a Model of Complexity Leadership. Paper presented at the International Conference on Complex Systems, Boston, 2006
73. Schreiber C, Carley K (2006) Leadership style as an enabler of organizational complex functioning. Emergence Complex Organ 8(4):61–76
74. Schreiber C, Marion R, Uhl-Bien M, Carley K (2007) Multi-Agent Based Modeling of Complexity Leadership Theory. Paper presented at the North American Association for Computational Social and Organizational Sciences, Atlanta, 2007
75. Schroeder M (1991) Fractals, chaos, power laws. Freeman, New York

76. Stephenson WD, Bonabeau E (2007) Expecting the Unexpected: The Need for a Networked Terrorism and Disaster Response Strategy (Electronic Version). Homeland Security Affairs, 3 http://www.hsaj.org/?article=3.1.3
77. Sterman JD (1989) Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. Manag Sci 35(3):321–339
78. Stringer SH (2006) Entertaining the Future. International Consumer Electronics Show Las Vegas January – 5:2006 Retrieved November, 2007, http://www.sony.com/SCA/speeches/060105_stringer.shtml
79. Taylor F (1911) The principles of scientific management. Harper, New York
80. Uhl-Bien M, Marion R (2005) Complexity leadership theory. Paper presented at the Workshop on Leadership and Complexity, George Washington University
81. Uhl-Bien M, Marion R (2007) The mechanisms of emergence in complexity leadership theory: A meso-model of adaptive dynamics in organizations. Paper presented at the Festschrift honoring Jerry Hunt, Lubbock, 2007
82. Uhl-Bien M, Marion R, McKelvey B (2007) Complexity Leadership Theory: Shifting Leadership from the Industrial Age to the Knowledge Era. Leadersh Q 18:298–318
83. Volberda HW (1996) Toward the flexible form: How to remain vital in hypercompetitive environments. Organ Sci 7(4):359–374
84. von Bertalanffy L (1956) General systems theory. In: L v. Bertalanffy, Rapoport A (eds) General systems: Yearbook of the society for the advancement of general systems theory, vol: 1. George Braziller Publishers, New York, pp 1–10
85. Weber M (1947) The theory of social and economic organization (trans: Henderson AH, Parsons T). Free Press, Glencoe

Social Phenomena Simulation

Paul Davidsson[1], Harko Verhagen[2]
[1] Blekinge Institute of Technology, Ronneby, Sweden
[2] Stockholm University and Royal Institute of Technology, Stockholm, Sweden

Article Outline

Glossary

Agent (or software agent) A self-contained entity that has a state and that is situated (able to perceive and act) in an environment. In addition, agents are often assumed to be rational and autonomous.

Cellular automaton A mathematical structure modeling a set of cells that interact with their neighbors. Each cell has a set of neighbors and a state. All the cells update their values simultaneously at discrete time steps. The new state of a cell is determined by the current state of its neighbors according to a local function or rule.

Microlevel simulation A type of simulation in which the specific behaviors of specific individuals are explicitly modeled.

Definition of the Subject

Social phenomena simulation in the area of agent-based modeling and simulation concerns the emulation of the individual behavior of a group of social entities, typically including their cognition, actions, and interaction. Agent-based social simulation constitutes the intersection of three scientific fields, namely, agent-based computing, the social sciences, and computer simulation [6]. Agent-based computing is a research area mainly within computer science and includes, e. g., agent-based modeling, design, and programming. By the social sciences we here refer to a large set of different sciences that study the interaction among social entities, e. g., social psychology, management science, policy, and some areas of biology. Computer simulation concerns the study of different techniques for simulating phenomena on a computer, e. g., discrete-event, object-oriented, and equation-based simulation.

Introduction

Computer simulation consists of three main steps: (i) designing a model of an actual or theoretical system, (ii) executing the model on a computer, and (iii) analyzing the execution output. Already in the early days of computer development, simulation was used in different research areas to predict the behavior of complex systems. Such simulations were typically based on differential equations and focused on results at the aggregate level. These models of, for instance, predator-prey populations could result in fairly accurate models but were limited in the sense that the models excluded individual behavior and decision making, as well as interaction between individuals, and were based on homogeneous agents. The development of agent-based modeling offers a possible solution to this problem with its (seemingly) natural mapping onto interacting individuals with incomplete information and capabilities, no global control, decentralized data, asynchronous computing, and inclusion of heterogeneous agents. Agent-based simulation models also offer the possibility of studying the dynamics of the interaction processes instead of focusing on the (static) results of these processes [16,26].

Agent-based modeling can be traced back to von Neumann, who in the 1950s invented what was later termed *cellular automata*. These were used by Conway in the 1970s when he constructed the well-known *Game of Life*. It is based on very simple rules determining the life and death of the cells in a virtual world in the form of a 2-D grid. Inspired by this work, researchers developed more-refined models, often modeling the social behavior of groups of animals or artificial creatures. One example is the Boid model by Reynolds [24], which simulates coordinated animal motion such as bird flocks and fish schools. With respect to human societies, Epstein and Axtell [8] developed in the 1990s one of the first agent-based models, called Sugarscape, to explore the role of social phenomena such as seasonal migrations, pollution, sexual reproduction, combat, and transmission of disease. This work is in spirit closely related to one of the best-known and earliest examples of the use of simulation in social science, namely, the Schelling model [27], in which cellular automata were used to simulate the emergence of segregation patterns in neighborhoods based on a few simple rules expressing the preferences of the agents. Another pioneer from the 1950s worth mentioning is Barricelli [2], who to some extent used agent-based modeling for simulating biological systems.

The cellular automata models closely resemble the models used in statistical physics, which has inspired physicists to include the simulation of social phenomena in large-scale social systems in their research agenda. In this area, sometimes referred to as sociophysics, phenomena such as opinion spreading in a society and competition between languages have been studied. These models originally described the behavior of atoms and molecules, which are quite simple objects, and the macrolevel phenomena caused by their interaction (rather than by complex behavior of the individual as in the case of humans). Thus, in these models little attention is paid to individual variation and the individual decision making is rather primitively modeled. A prominent example of sociophysics is the work of Galam [10].

To sum up, we can identify two main approaches to social simulation:

- Macrolevel (or equation-based) simulation, which is typically based on mathematical models. It views the set of individuals (the population) as a structure that can be characterized by a number of variables.
- Microlevel (or agent-based) simulation, in which the specific behaviors of specific individuals are explicitly modeled. In contrast to macrolevel simulation, it views the structure as emerging from the interactions between individuals and thus exploring the standpoint that complex effects need not have complex causes.

As argued by Van Parunak et al. [21], agent-based modeling is most appropriate for domains characterized by a high degree of localization and distribution and dominated by discrete decision. Equation-based modeling, on the other hand, is most naturally applied to systems that can be modeled centrally and in which the dynamics are dominated by physical laws rather than information processing. We will here focus on agent-based models, particularly those that have a richer representation of the individual than the cellular automata and statistical physics models.

Why Simulate Social Phenomena?

Simulation of social phenomena can be done for different purposes, e. g.,

- Supporting social-theory building;
- Supporting the engineering of systems, e. g., validation, testing, etc.;
- Supporting planning, policy making, and other decision making;
- Training, in order to improve a person's skills in a certain domain.

It is possible to distinguish between four types of end users: *scientists*, who use social phenomena simulation in the research process to gain new knowledge, *policymakers*, who use it for making strategic decisions, *managers* (of systems), who use it to make operational decisions, and *other professionals*, such as architects, who use it in their daily work. We will now describe how these types of end users may use simulation of social phenomena for different purposes.

Supporting Social-Theory Building

In the context of social-theory building, agent-based simulation can be seen as an experimental method or as theories in themselves [26]. In the former case, simulations are run to test the predictions of theories, whereas in the latter case simulations in themselves are formal models of theories. Formalizing the ambiguous, natural-language-based theories of the social sciences helps to find inconsistencies and other problems, and thus contributes to theory building.

Using agent-based simulation studies as an experimental tool offers great possibilities. Many experiments with human societies are either unethical or even impossible to conduct. Experiments in silico, on the other hand, are fully possible. These can also breathe new life into the ever-present debate in sociology on the micro–macro link [1]. Agent-based models mostly focus on the emergence of macrolevel properties from the local interaction of adaptive agents that influence one another [17,26]. However, simulations in computational organization theory [4,22], for example, often try to analyze the influence of macrolevel phenomena on individuals. Using agent-based models to simulate the bidirectional relation between micro- and macrolevel concepts would provide tools to analyze the theoretical consequences of the work done by theorists such as Habermas, Giddens, and Bourdieu, to name a few [26].

Supporting the Engineering of Systems

Many new technical systems are distributed and involve complex interactions between humans and machines. The properties of agent-based simulation make it especially suitable for simulating these kinds of systems. The idea is to model the behavior of human users in terms of software agents. In particular, this seems useful in situations where it is too expensive, difficult, inconvenient, tiresome, or even impossible for real human users to test out a new technical system. Of course, also the technical system, or parts thereof, may be simulated. For instance, if the technical system includes hardware that is expensive and/or special purpose, it is natural to simulate also this part of the system when testing out the control software. An example of such a case is the testing of control systems for "intelligent buildings," where agents simulate the behavior of the people in the building [5].

Supporting Planning, Policy Making, and Other Decision Making

Here the focus is on exploring different possible future scenarios in order to choose between alternative actions. Besides this type of prediction, simulation of social phenomena may be used for analysis, i. e., to gain deeper knowledge and understanding of a certain phenomenon.

An area in which several studies of this kind have been carried out is disaster management, such as experiments concerning different roles and the efficiency of reactions to emergencies [18]. Based on individuals' observations, personal characteristics and skills, past experience and role characteristics, and social network, the agents create a plan to execute. Each agent represents a human being (acting in a particular role). The effect of adding a role (floor warden) in a fire alarm scenario upon the evacuation efficiency in an abstract environment is analyzed. In another approach, the agents are placed in an environment based on GIS (geographical information system) data, thereby tying the simulation closer to the physical reality [29]. In yet another study, real-world data were used for both the environment and the agents' internal decision-making model to analyze the effect of different insurance policies on the willingness of agents to pay for a disaster insurance policy [3].

Another application area for this type of simulation study is disease spreading. Typically, agents are used to represent human beings and the simulation model is linked to real-world geographical data. One study [32] also included agents that represent towns acting as the epicenter of disease outbreak. The town agent's behavior repertoire consisted of different containment strategies. The simulation model can be quickly adapted to local circumstances via the geographical data (given that there is data on the population as well) and is used to determine the effects of different containment strategies.

A third area where agent-based social simulation has been used to support planning and policy making is traffic and transport. An example of this is the simulation of all car travel in Switzerland during morning peak traffic [23].

Training

The main advantage of using simulation for training purposes is to be part of a real-world-like situation without real-world consequences. Especially in the military the use of simulation for training purposes is widespread. Also in medicine, where mistakes can be very expensive in terms of money and lives, the use of simulation in education is on the rise.

An early product in this area was a tool to help train police officers to manage large public gatherings such as crowds, demonstrations, and marches [31]. Another early example of agent-based simulation for training purposes is Steve [19,25]. Steve was an agent integrated with voice synthesis software and virtual reality software providing a very realistic training environment for controlling the engine room of a virtual US Navy surface ship.

An example of a more recent project is the PSI agent [15]. Whereas in most cases the simulator training is aimed at training practical skills or decision making, this work focuses on acquiring theoretical insights in the realm of psychological theory. The simulation enables students to explore psychological processes without ethical problems.

Simulating Social Phenomena

One of the first, and most simple, way of performing microlevel simulation is often called *dynamic microsimulation* [11,12]. It is used to simulate the effect of the passing of time on individuals. Data from a (preferably large) random sample from the population to be simulated is used to initially characterize the simulated individuals. Some examples of sampled features are: age, sex, employment status, income, and health status. A set of transition probabilities are used to describe how these features will change over a given time period, e. g., there is a probability that an employed person will become unemployed over the course of a year. The transition probabilities are applied to the population for each individual in turn and then repeatedly reapplied for a number of simulated time periods. Sometimes it is necessary to also model changes in the population, e. g., birth, death, and marriage. This type of simulation can be used to, e. g., predict the outcome of different social policies. However, the quality of such simulations depends on the quality of:

- the random sample, which must be representative, and
- the transition probabilities, which must be valid and complete.

In traditional microsimulation, the behavior of each individual is regarded as a "black box." The behavior is modeled in terms of probabilities and no attempt is made to justify these in terms of individual preferences, decisions, plans, etc. Also, each simulated individual is considered in isolation without regard to their interaction with others. Thus, better results may be gained if cognitive processes and communication between individuals are also simulated.

Opening the black box of individual decision making can be done in several ways. The first layer to add is often individual psychology; for instance, the so-called beliefs, desires, and intentions (BDI) model is often used. Models of individual cognition used in agent-based social simulation include the use of Soar (a computer implementation of Allen Newell's unified theory of cognition [20]), which was used in Steve (discussed in Sect. "Why Simulate Social Phenomena?").

For the simulation of social behavior the agents need to be equipped with mechanisms for reasoning at the social level (unless the social level is regarded as emerging from individual behavior and decision making). Several models have been based on theories from economics, social psychology, sociology, etc. An example of this is provided by Guye-Vuillème [13], who has developed an agent-based model for simulating human interaction in a virtual-reality environment. The model is based on sociological concepts such as roles, values, and norms and motivational theories from social psychology to simulate persons with social identities and relationships. Another example is the Consumat model [14], a metamodel combining several psychological theories on decision making in a consumer situation, used, for instance, to investigate different flood-management policies [3]. Also, nonsymbolic approaches such as neural networks have been used to model agents' decision making [18].

Future Directions

In a recent study of applications of agent-based simulation [7], it was concluded that even if agent-based simulation seems a promising approach to many problems involving the simulation of complex systems of interacting entities such as social phenomena, it seems that the full potential of the agent concept often is not utilized. For instance, most models have very primitive agent cognition, in particular if the number of agents involved is large.

Regarding future applications, Fiedrich and Burghardt [9] argue that agent-based simulation is a very promising approach to disaster management practice. In particular, agent-based social simulation in combination with sophisticated visualization techniques, such as virtual reality, in the form of "serious games," has the potential to provide very powerful training environments. In the context of military training, Stone [28] provides some interesting applications.

Further Reading

Classic books in the area of simulation of social behavior include "Growing Artificial Societies: Social Science from the Bottom Up" by Epstein and Axtell [8] and "Simulation for the Social Scientist" by Gilbert and Troitzsch [12].

More recent findings can be found in, e. g., the Journal of Artificial Societies and Social Simulation (http://jasss.soc.surrey.ac.uk) and the proceedings of, e. g., the International Workshop series on Multi-Agent-Based Simulation (MABS) (http://www.pcs.usp.br/~mabs/), the World Congress in Social Simulation (WCSS) [30], the conference of the European Social Simulation Association (ESSA) (http://www.essa.eu.org/), and the series of Agent Workshops in Chicago (http://www.agent2005.anl.gov/).

Bibliography

1. Alexander JC, Giesen B, Münch R, Smelser NJ (eds) (1987) The Micro-Macro Link. University of California Press, Berkeley
2. Barricelli NA (1957) Symbiogenetic evolution processes realized by artificial methods. Methodos 9(35–36):143–182

3. Brouwers L, Verhagen H (2004) Applying the Consumat model to flood management policies. In: Müller J, Seidel M (eds) 4th Workshop on Agent-Based Simulation. SCS, Montpellier, pp 29–34
4. Carley KM, Prietula M (eds) (1994) Computational Organization Theory. Erlbaum, Hillsdale
5. Davidsson P (2000) Multi-agent-based simulation: beyond social simulation. In: Moss S, Davidsson P (eds) Multi-Agent-Based Simulation. Lecture Notes in Computer Science, vol 1979. Springer, Berlin, pp 98–107
6. Davidsson P (2002) Agent-based social simulation: a computer science view. J Artif Soc Soc Simul 5(1)
7. Davidsson P, Holmgren J, Kyhlbäck H, Mengistu D, Persson M (2007) Applications of multi-agent-based simulation. In: Antunes L, Takadama K (eds) Multi-Agent-Based Simulation VII. Lecture Notes in Computer Science, vol 4442. Springer, Berlin
8. Epstein JM, Axtell RL (1996) Growing Artificial Societies: Social Science from the Bottom Up. MIT Press, Cambridge
9. Fiedrich F, Burghardt P (2007) Emergency response information systems: emerging trends and technologies: agent-based systems for disaster management. Commun ACM 50(3):41–42
10. Galam S (2008) Sociophysics: a review of Galam models. arXiv:0803.1800 http://arxiv.org/abs/0803.1800. Accessed 06 Aug 2008
11. Gilbert N (1994) Computer Simulation of Social Processes. Social Research Update, Issue 6, Department of Sociology, University of Surrey, UK. http://www.soc.surrey.ac.uk/sru/. Accessed 06 Aug 2008
12. Gilbert N, Troitzsch KG (2005) Simulation for the Social Scientist, 2nd edn. Open University Press, Maidenhead
13. Guye-Vuillème A (2004) Simulation of nonverbal social interaction and small groups dynamics in virtual environments. Ph D thesis, Ècole Polytechnique Fédérale de Lausanne, No 2933
14. Janssen MA, Jager W (1999) An Integrated Approach to Simulating Behavioural Processes: A Case Study of the Lock-in of Consumption Patterns. J Artif Soc Soc Simul 2(2)
15. Künzel J, Hämmer V (2006) Simulation in university education: the artificial agent PSI as a teaching tool. Simulation 82(11):761–768
16. Lansing JS (2002) “Artificial societies” and the social sciences. Artif Life 8:279–292
17. Macy MW, Willer R (2002) From factors to actors: computational sociology and agent-based modeling. Annu Rev Sociol 28:143–166
18. Massaguer D, Balasubramanian V, Mehrotra S, Venkatasubramanian N (2006) Multi-agent simulation of disaster response. In: Proceedings of the 1st International Workshop on Agent Technology for Disaster Management 2006. Hakodate
19. Méndez G, Rickel J, de Antonio A (2003) Steve meets Jack: the integration of an intelligent tutor and a virtual environment with planning capabilities. In: Intelligent Virtual Agents. Lecture Notes on Artificial Intelligence, vol 2792. Springer, Berlin, pp 325–332
20. Newell A (1994) Unified Theories of Cognition. Harvard University Press, Cambridge
21. Parunak HVD, Savit R, Riolo RL (1998) Agent-based modeling vs. equation-based modeling: a case study and users’ guide. In: Sichman JS, Conte R, Gilbert N (eds) Multi-Agent Systems and Agent-Based Simulation. Lecture Notes in Computer Science, vol 1534. Springer, Berlin, pp 10–26
22. Prietula MJ, Carley KM, Gasser L (eds) (1998) Simulating Organizations: Computational Models of Institutions and Groups. MIT Press, Cambridge
23. Raney B, Cetin N, Völlmy A, Vrtic M, Axhausen K, Nagel K (2003) An Agent-Based Microsimulation Model of Swiss Travel: First Results. J Netw Spatial Econ 3(1):23–41
24. Reynolds CW (1987) Flocks, herds, and schools: a distributed behavioral model. Comput Graph 21(4):25–34 (SIGGRAPH ’87 Conference Proceedings)
25. Rickel J, Lewis Johnson W (1999) Animated agents for procedural training in virtual reality: perception, cognition, and motor control. Appl Artif Intell 13:343–382
26. Sawyer RK (2003) Artificial societies – multi-agent systems and the micro-macro link in sociological theory. Sociol Meth Res 31(3):325–363
27. Schelling TC (1971) Dynamic models of segregation. J Math Sociol 1(1):143–186
28. Stone B (2005) Serious gaming. Def Manag J (31):142–144
29. Takahashi T (2007) Agent based disaster simulation evaluation and its probability model interpretation. In: Proceedings of the 4th International Conference on Intelligent Human-Computer Systems for Crisis Response and Management, 2007, Deltft, pp 369–376
30. Takahashi S, Sallach D, Rouchier J (2007) Advancing Social Simulation – The First World Congress. Springer, Berlin
31. Williams R (1993) An agent based simulation environment for public order management training. In: Western Simulation Multiconference, Object-Oriented Simulation Conference (OOS ’93), San Diego 151–156
32. Yergens D, Hiner J, Denzinger J, Noseworthy T (2006) IDESS – A multi-agent-based simulation system for rapid development of infectious disease models. Int Trans Syst Sci Appl 1(1):51–58

Social Processes, Physical Models of

FRANTIŠEK SLANINA[1,2]

[1] Institute of Physics, Academy of Sciences of the Czech Republic, Prague, Czech Republic

[2] Center for Theoretical Study, Prague, Czech Republic

Article Outline

Glossary

Absorbing state In a stochastic process, a state from which there is no way out. Once the system reaches the absorbing state, it stays there forever.

Complete graph A graph where every pair of vertices is connected by an edge.

Configuration space The ensemble of all allowed states (configurations) a model system can reach. Any point in configuration space represents one such state.

Graph an assembly of points (vertices), some of which are connected by lines (edges).

Hypercubic lattice A generalization of a rectangular plane mesh; a graph embedded in a space of arbitrary dimensionality d. Assuming rectangular coordinates $x_1, x_2, \ldots, x_d$, the coordinates of the vertices in a hypercubic lattice are all whole numbers.

Markov process A special type of stochastic process. In a Markov process the system changes its state randomly and, most importantly, the changes of state are independent of history. A Markov process is a memoryless system.

Random walk A stochastic process describing hops of a particle to randomly chosen neighbor sites on a lattice.

Stochastic process A sequence of random events. At each time t, which may be either discrete ($t = 1, 2$, etc.) or continuous, a new random variable is introduced representing the outcome of the process in that time. Also called a random process.

Definition of the Subject

Modeling social phenomena as if they were manifestations of mutual interactions of physical objects is the ultimate goal of the reductionist approach to reality. Both the inanimate and animate worlds, including all the behavior of humans, would be traced back to the properties of atoms and molecules. This program is absolutely unrealizable, though. On the other hand, the discipline of **sociophysics** tries to bypass the brute-force approach by developing schematically effective models which aim at describing reality at a "macroscopic", rather than microscopic, level.

For example, when one wants to model the behavior of a large assembly of humans facing the necessity of choosing between two options, it is customary to neglect all details of the behavior of the people involved and describe their states by two-value quantities, such as $s = +1$ or $s = -1$; physicists call them spins. The interactions are often expressed using a cost function, which physicists call energy. The state with lowest energy is favored, but there are also external perturbations, or noise, which prevent the system from settling in that state. Physicists call the measure of the noise the temperature. The system of interacting spins at a certain temperature then serves as a model of the particular situation in human society.

It is a non-trivial question if such approach could work and if so, why. The essential ingredient for its functioning is **universality**. This notion, borrowed from statistical physics, means, that the behavior of the system does not ultimately depend on the details of that system, but only on its generic features. In plain words, for the description of the human mind, the behavior of a driver on a highway, or of a stockbroker, it is irrelevant whether the matter is composed of quarks, strings or whatever elementary particles may be discovered. The description of the world at the "macroscopic" level is detached from "microscopic" details.

It is the task of **sociophysics** to develop suitable macroscopic models that conform to observed facts about society and, at the same time, yield themselves to the methods of theoretical physics. There are non-negligible successes in this effort, yet still many areas remain unexplored. Consequently, the subject is alive and in a state of rapid progress.

Historical Introduction

It has been nearly two hundred years since the notion of social physics was introduced by the French philosopher Auguste Comte [1] in his attempt to revolutionize all thinking under the banner of positive philosophy. He himself abandoned that term later [2], but the idea of connecting social and physical phenomena in a unified frame has been around since. Later on, it somehow degenerated into a purely descriptive science, borrowing from physics nothing more than the tendency to exactitude in measurements; a prominent name to cite in this context is the Belgian statistician Adolphe Quételet. Little use stems from twopenny analogies, advocated by some later followers of Hegelian philosophy, between phase transitions, such as boiling water, and abrupt changes, mostly violent ones, in human society. Nonetheless, you may still find these considerations, based on the "quantity turns into quality" principle, in popular-philosophy brochures.

Skipping ahead a little more than a century and half since Comte, the first modern attempt to bridge the gap between social sciences and physics was marked by a memorable conference scheduled to take place in Moscow, 1 to 5 July 1974. Scientists both from the West and from the USSR were to discuss implications of physics for other fields, including social sciences and humanities. The organizing committee included people such as Kenneth Arrow, Nobel laureate for Economy and Hans

Bethe, Nobel laureate for Physics. Among the speakers was Andrei Sakharov, the well-known dissident and Nobel laureate for Peace. However, the Communist authorities stepped in, the conference was banned, most of the Russian participants were arrested, and a majority of them were exiled. However, many of the manuscripts were smuggled from the USSR to the West, and eventually the contributions were published in a proceedings volume [3]. A small sample of it appeared in [4].

In the 1980s, a very fruitful and general concept of synergetics was developed by H. Haken and his followers. The main idea consists in spontaneous emergence of coherent structures. In addition to being illustrated by examples from several branches of physical sciences (e. g. lasers) the idea was also elaborated in the context of social phenomena [5].

About the same time, the science of complexity started to gain wider popularity (and continues to do so to this day). The Santa Fe Institute in Santa Fe, USA, pioneered this research; it was here that complex system theory began to be systematically applied to economics and, more generally, to all human social behavior. Another Physics Nobel laureate, P. W. Anderson [6], has played a prominent role in promoting this path.

To finish this historical overview, the journey returns to France, where since the early 1980s the group of Serge Galam has been busy developing genuine physical models of social phenomena, also using he term sociophysics consistently [7,8,9]. Let us stop here and turn to several selected topical problems, where physically based models may help answer social scientists' questions.

Individualism Versus Cooperation

People never earn their living alone. From the darkest prehistorical past they have joined together to form gangs, stalking and hunting their prey. Without communication within the band, our ancestors could hardly have caught a mammoth, language would never have evolved and human brain capacity would probably never have exceeded that of a lemur. In short, cooperation between humans has been decisive in shaping the world around us. A tendency to cooperation is an inherent feature of human nature.

It was not until the middle of the eighteenth century or so that a different view started to spread. In this view, people are presented as selfish profit-seekers and if any cooperation is observed, it occurs *despite* the natural tendency to maximize personal gain, or as a by-product of that tendency. Under vague names such as neo-liberalism and social Darwinism, these and related ideas pervade the current thinking on human society.

On the other hand, empirically, cooperation is much more evolved now and assumes more sophisticated and complex forms than in the old times when it was considered self-evident. To detect this more complex cooperation, however strong it may be, requires external explanation. The starting point is a zero hypothesis of non-cooperation, from which cooperation emerges by some nontrivial mechanism to be discovered.

Prisoner's Dilemma Game

Games in General *Game theory* [10,11], introduced in 1940s by John von Neumann and Oskar Morgenstern provides very fertile ground on which to model the starting assumption of selfish individuals. In this model, agents are not only utterly selfish, but also absolutely rational. In a typical setup, agents meet in pairs and each of them chooses one of S possible strategies. If the first agent adopts strategy i and the second agent strategy j, then the first one gains an amount denoted A_{ij}, while the other's gain is A_{ji}. The matrix A is called a *payoff matrix*, and various types of games are distinguished according to its properties. For example, if the gain of one agent equals the loss of the other agent, so $A_{ij} + A_{ji} = 0$ for all i and j, it is a zero-sum game. If the gain of the winning party is always smaller than the loss of the adversary, the game is a negative-sum one, so it holds that $A_{ij} + A_{ji} < 0$ for all i and j.

It is assumed that the rules of the game, quantified in the matrix A, are known to both players. Therefore, they can build their strategies on rational analysis of the payoff matrix. It may happen that one strategy gives highest gain, irrespective of the action taken by the opponent. Formally, the first player's strategy, k, is such that $A_{ij} \leq A_{kj}$ for any strategy i of the first player and any strategy j of the second player. The same may also hold for the second player. Suppose that l is her best strategy, irrespective of the action of the first player. Obviously then, the first person always plays k while the second always plays l. If either player changes his or her strategy unilaterally, he or she is instantly worse off. Such a situation, if it happens, is called Nash equilibrium, after John F. Nash, a mathematician who devoted much of his career to applications of game theory in economics [12,13].

It is vital to understand that the notion of Nash equilibrium differs fundamentally from the usual equilibrium studied in various branches of physics. Whether the problem is to find the equilibrium of solid bodies on a lever or thermodynamic equilibrium in a system of steam, water and ice, the situation is formalized by finding a unique function to be minimized (energy, for example). On the contrary, Nash equilibrium means that every player maximizes his or her own function, with state of all other players fixed [14].

Introduction to Prisoner's Dilemma Collaboration has been studied using a very simple two-player game with two strategies, called the prisoner's dilemma game. Imagine that the police have arrested two accomplices for a burglary, but do not have enough evidence to prove that they actually committed the crime. The prisoners are kept well separated and the investigator offers a deal to each of them independently. Each criminal is given a promise that if he parts company with the other and confesses to the robbery they did together, he will be rewarded. If both of them confess, they will be jailed only for a short period; if one confesses and the other does not, then the first one is released and rewarded and the other gets a severe punishment; if neither of them confesses, both are released. Each must choose without knowing the other's choice.

The strategy of collaboration (C) dictates not to confess. In fact, it would be most beneficial to both suspects, if both collaborate. However, individually it is more tempting to defect (D) and confess to the police, as it prevents the situation in which the other confesses and the individual is punished alone. If both prisoners reckon in this way, the result is that both defect and must suffer some time in prison. Hence the dilemma.

The payoff matrix of this prisoner's dilemma game is characterized by four numbers, the gain if both defect, $A_{DD} = P$, or both cooperate, $A_{CC} = R$, and the gain of the defector $A_{DC} = T$ and the collaborator $A_{CD} = S$, if one of them defects and the other does not. So,

$$A_{PD} = \begin{array}{cc} C\,D & \\ \begin{pmatrix} R\,S \\ T\,P \end{pmatrix} & \begin{matrix} C \\ D \end{matrix} \end{array} . \tag{1}$$

In order for the prisoner's dilemma to work as described above, the values must satisfy the inequalities $T > R > P > S$ and $2R > S + T$. The most studied values of the parameters are

$$T = 5\,, \quad R = 3\,, \quad P = 1\,, \quad S = 0\,. \tag{2}$$

It is easy to see that there is a single Nash equilibrium: the case in which both players defect. We arrived at this conclusion intuitively, above, and an exact check is even quicker. If the second player collaborates, the rewards of the first are R and T if he collaborates or defects, respectively; because $T > R$, defection is better. Similarly, if the second player defects, the gains for the first are S and P, and as $P > S$, it is again preferable to defect. So, defection is the optimal strategy.

Evolution of Cooperation

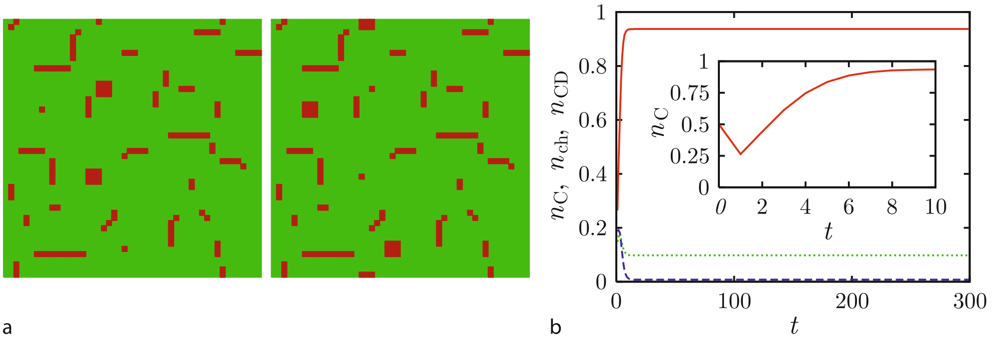
Iterated Prisoner's Dilemma The prisoner's dilemma game is rather trivial: if the players meet only once, defecting is certainly the best strategy. The Nash equilibrium excludes cooperation a priori, coinciding with the neo-liberalist view of society. What, then, is the mechanism behind the cooperation seen in reality? If there is a way for emergence of cooperation in the prisoner's dilemma game, perhaps we can see why reality sometimes follows that path.

Things become much more complex if the players face each other in an unlimited series of encounters. Robert Axelrod showed [15,16,17], in a series of computer tournaments, that complex strategies, taking into account past actions of the adversary, perform much better than simple defection. The generic and repeatedly confirmed outcome was that the strategy called Tit-For-Tat (TFT) outperformed all the rest of the strategies tried. It consists in playing cooperation at first and then repeating the last action of the other player.

Even more complicated behavior emerges if there is some level of noise. The players can make mistakes, or sometimes play at random, not obeying their usual strategies. New strategies may emerge due to mutations. Also relevant is the length of the players' memories, i. e. how many steps in the past they remember their opponents' actions. When all agents are put into a common room, where all of them they can play with all others, mixtures of strategies emerge [18,19]. The composition of the mixture depends on the length of the memory of the agents; if the memory length is unlimited, the evolution of the system proceeds indefinitely, introducing more and more complex strategies all the time.

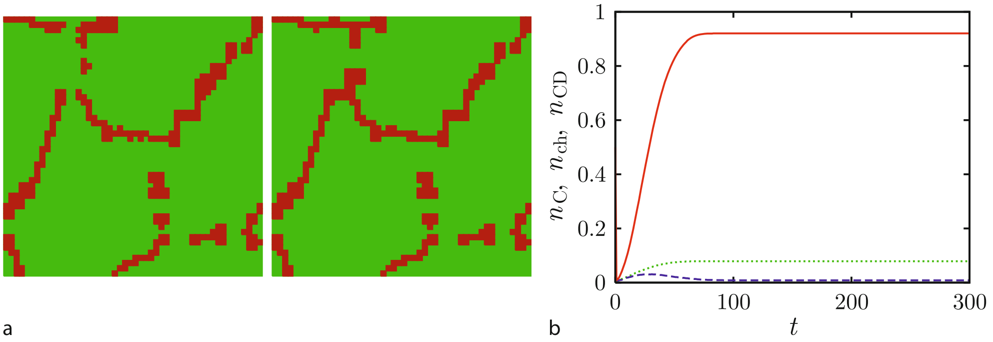
Spatial Prisoner's Dilemma The trivial prisoner's dilemma game becomes complex when the players play again and again with the same opponent. But, extending the model to cover a wider field, how it could be that among billions of people a player finds the same opponent he or she has already faced? If the two were molecules in a well-stirred container, their chance to meet repeatedly would be virtually nil. Some *compartmentalization* of the agents must be imposed, in order to apply the iterative prisoner's dilemma to a large number of players. But the influence of compartmentalization can also be studied independently, preferably in its extreme version, where immobile agents are fixed at the vertices of a network, for example a two-dimensional square lattice: that is the structure used throughout this subsection.

Agents play the usual prisoner's dilemma game with their eight neighbors. The complexity stems from their



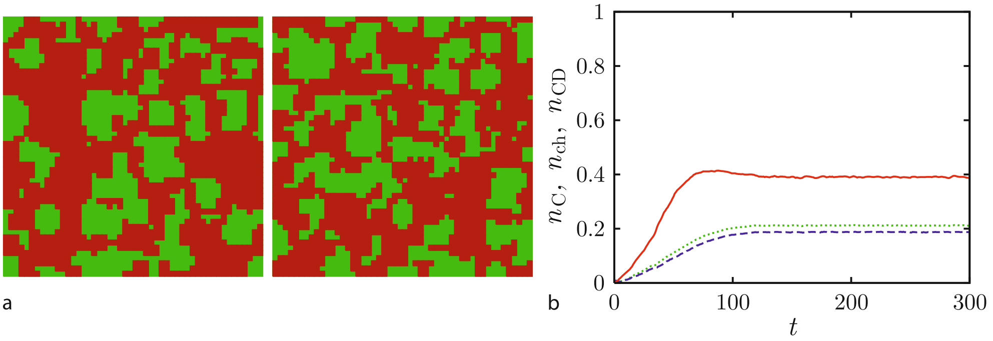
Social Processes, Physical Models of, Figure 1

Spatial prisoner's dilemma. In the *left two panels*, the configuration is drawn in times t_1, t_2, $t_1 < t_2$, on the lattice with $L = 50$. Cooperating agents are shown in *green*, the defectors in *red*. In the *right panel*, time evolution of the density of cooperators n_C (*solid line*), density of agents changing their state from one step to the next n_ch (*dashed line*), and density of interfaces n_CD (*dotted line*) are shown. System size is $L = 100$, data are averaged over 100 realizations. The incentive to defect is, $b = 1.05$. The initial condition is random spatial distribution of cooperators with density $n_\mathrm{C} = 0.5$. In the *inset*, the detail for short times is shown



Social Processes, Physical Models of, Figure 2

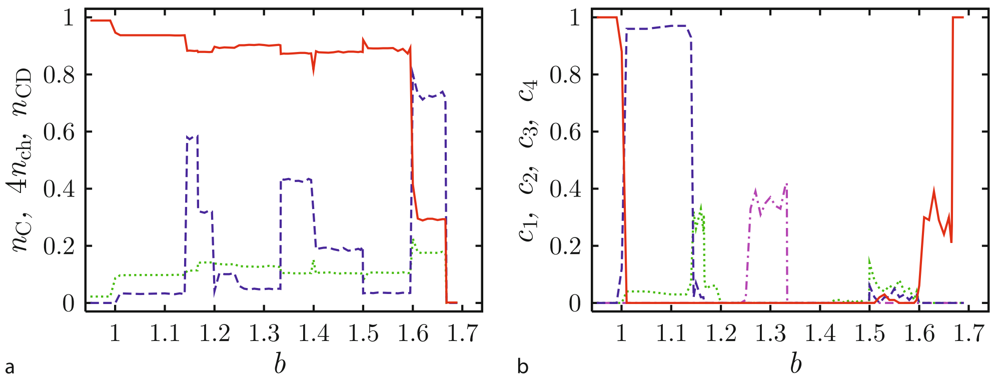
Same as Fig. 1 with $b = 1.5$



Social Processes, Physical Models of, Figure 3

Same as Fig. 1 with $b = 1.6$

ability to change their strategies from one step to the next [20]. For simplicity, let us consider only the 0-memory strategies: collaboration and defection. At each round, the agent adds her gain from the eight plays with her neighbors, who do the same in their turn. Each agent then decides upon her strategy for the next step. She looks at the gains of her neighbors and compares these numbers with her own. If any of the neighbors earns more than she, the agent adopts the current strategy of the most successful neighbor for the next step. Otherwise, the agent repeats her previous strategy. The evolution of the strategies proceeds by imitation.



Social Processes, Physical Models of, Figure 4

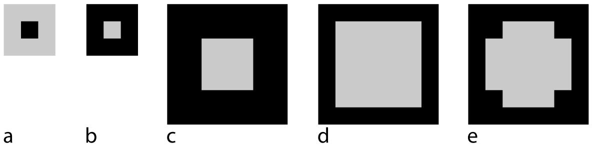
Spatial prisoner's dilemma. Various parameters of the stationary state depending on the incentive to defect *b*. All data are calculated for a lattice with $L = 100$; initial configuration is a random spatial distribution of cooperators with density $n_C = 0.5$; data are averaged over 100 realizations. The *left panel* shows the density of cooperators (*solid line*), density of sites changing their state from one step to the next (*dashed line*), and density of bonds connecting cooperators with defectors (*dotted line*). The *right panel* shows the fraction of realizations reaching static configuration (*solid line*), and fractions reaching cyclic attractors with period 2 (*dashed line*), 3 (*dot-dashed line*), and 4 (*dotted line*)

The payoff matrix can be simplified so that there is only one control parameter, the temptation to defect b. It is assumed that $T = b > 1$, $R = 1$, and $P = S = 0$. The spatial prisoner's dilemma game is then simulated, starting from a random configuration, where every agent chooses cooperation or defection with equal probability. Further evolution proceeds by deterministic parallel dynamics, as described above. In fact, the system is a cellular automaton with specific, relatively complicated, update rules. The resulting configurations are illustrated in Figs. 1 to 3. The cooperators survive in significant proportion even if the temptation for defection b is large. However, their spatial arrangement strongly depends on the value of b. Figure 1 shows the situation for the value $b = 1.05$, only slightly above the lowest limit compatible with prisoner's dilemma inequalities. Small groups of defectors survive within a large sea of cooperators. Most of the defector groups are stable or change cyclically with a short period. When the temptation is increased to $b = 1.5$, isolated islands of defectors grow into strings joined together at some places, as seen in Fig. 2. Cooperation still prevails, but only within patches encircled by defectors. The spatial structure exhibits only small variations in time. This changes when the temptation grows further. Figure 3, where $b = 1.6$, already shows more defectors than collaborators and the arrangement changes chaotically.

Systematic study of the dependence on b is summarized in Fig. 4 showing data from simulations, averaged over many realizations of the initial conditions. First, the fraction of cooperators, n_C, remains quite high, above 80%, up to about $b = 1.6$, where it drops suddenly to about 40%; only above $b = 1.7$ do the cooperators vanish. Clearly, repeated games induced by fixed spatial arrangement of the players strongly encourages cooperation. Note that the players do not follow any strategy based on observation of the past behavior of the agents. The cooperation emerges spontaneously.

Further inspection of the dependence of n_C on b reveals sudden jumps at specific values of the temptation parameter. The jumps are even more pronounced in some other parameters characterizing the stationary state, namely the concentration n_{ch} of sites changing their state from one step to the next, and the concentration of bonds (neighbor pairs) connecting a defector and a cooperator, n_{CD}, which is the measure of the density of interfaces between cooperating and defecting domains. Furthermore, some realizations end in static configurations, while other reach a cyclic attractor with short periods 2, 3, or 4. The fraction of realizations corresponding to these four types are denoted c_1, c_2, c_3, and c_4, respectively. Of course, some initial conditions may also lead to longer stationary periods or to a quasi-chaotic state with a barely identifiable periodicity. Interestingly, the quantities c_i depend on b in a very irregular fashion. For example, the period-2 states dominate in an interval from $b = 1$ to about $b = 1.14$, but are rare elsewhere. In contrast, the period-3 states occur practically only in the interval $5/4 < b < 4/3$.

To understand these features it is necessary to analyze various spatial structures produced in the dynamics [21]. For example, an isolated defector in the sea of cooperators, as shown in Fig. 5a, has 8 cooperating neighbors, so its gain is $g_D = 8p$. The neighbors themselves have 7 cooperating neighbors, resulting in gain $g_C = 7$. This means that for $p < 7/8$ the defector becomes cooperator in the next step,



Social Processes, Physical Models of, Figure 5
Several configurations of the defectors (*black*) and cooperators (*gray*) in the spatial prisoner's dilemma game on a square lattice. Their stability is discussed in the text

but for $p \geq 7/8$ it survives. Similar analysis shows that an isolated cooperator surrounded by defectors, Fig. 5b, is never stable. Indeed, the cooperator gains 0, but each of the the neighboring defectors gains p.

It is possible to proceed further to more and more complicated geometries. For example, a 2×2 square of cooperators, Fig. 5c, is much more stable than an isolated cooperator. Each of the four gains $g_C = 3$ from its three cooperating neighbors. The defecting neighbors are of two types. At the corners, they have one cooperator to exploit, so their gain is p. The defectors adjacent to the edges of the square have two cooperating neighbors, giving higher gain $g_C = 2p$. Therefore, the square persists for $p \leq 3/2$. But that is not all; the defecting neighbors see that the cooperators gain more, so they become cooperators themselves in the next step. For $p < 3/2$ the initial 2×2 square grows into a 4×4 square of cooperators, Fig. 5d, which in turn expands into 6×6 square, and so on. Cooperation spreads despite the relatively large value of the temptation to defect!

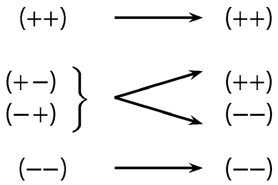
Some configurations may exhibit periodic changes. A multitude of possible generalizations can be found in the literature, either concerning update rules [22,23,24,25] or geometry of the links connecting interacting neighbors [26,27,28,29,30]. However, the general features of the spatial prisoner's dilemma game remain in force, namely the fact that repeated plays against the same agents, which are dictated by the geometry of the social network, leads naturally to coexistence of large patches of collaborators alternated with various arrangements of defectors.

To sum up, spontaneous emergence of cooperation seems to be reproduced in model situations, on condition that the game is played repeatedly. The overall picture of cooperation is typically rather complex, precluding any simplistic ideology-based conclusions.

Opinion Dynamics

Voter Model

People can make up their minds by simply looking around and picking the opinions of randomly chosen neighbors.



Social Processes, Physical Models of, Figure 6
Illustration of the dynamics of opinions in the voter model. The scheme shows how pairs of neighboring sites are updated. If the two sites have identical states, they remain unchanged. If the single-site states differ, they can become either both + or both − with equal probability. Conservation of average magnetization follows directly

That is the idea behind the stochastic process introduced in 1970s [31,32] and called the *voter model*. The voter model plays a special role among other models of opinion spreading and consensus formation, because it is exactly soluble in any spatial dimension, while showing highly non-trivial dynamics [33,34,35,36,37,38,39,40,41]. Physicists are interested in the voter model, as it can shed light on spinodal decomposition [35] and chemists use it to model catalytic reactions [36,37,42].

Definition In fact, there is a whole family of diverse voter models [43]; the exactly soluble class comprises the so-called linear voter models.

The most studied geometry is the d-dimensional hypercubic lattice Λ of linear dimension L, with periodic boundary conditions. The coordinates of the point $x \in \Lambda$ will be denoted $x_\alpha, \alpha = 1, 2, \ldots, d$.

On each lattice site there is an agent whose state can be either $+1$ or -1. These two choices can represent a person's political preferences in a two-party system, hence the name voter model. The configuration of the entire system is described by a point in the configuration space $\sigma \in S = \{-1, +1\}^\Lambda$. The state of the site $x \in \Lambda$ is denoted $\sigma(x)$.

The dynamics of the model are very simple. In each step, one site x and its neighbor y are chosen randomly. Then x adopts the state of y, so $\sigma(x)$ is replaced by $\sigma(y)$, as illustrated in Fig. 6. This scheme also demonstrates one important property of the voter model. The average opinion, which a physicist would call magnetization, defined as $m = (1/|\Lambda|) \sum_{x\in\Lambda} \sigma(x)$ is conserved after averaging all possible realizations of the process, even though in individual realizations it may fluctuate. It is also evident that the uniform states where all sites are either $+1$ or -1 never change. These two configurations are *absorbing states* of the voter model.

In one dimension the dynamics can be easily understood. The configuration is determined by the sequence of "domain walls", separating regions uniformly populated by $+1$ or -1. The configuration can change only by flipping the state of the sites beside the domain wall. Either the agent to the left of the wall adopts the state of the agent on the right side, or vice versa. In the former case the domain wall moves one step leftwards, in the latter case it jumps rightwards. Both possibilities have the same probability, so the domain wall performs a random walk. Moreover, when two domain walls meet, the region bordered by them disappears and the domain walls themselves annihilate. So, the one-dimensional voter model is exactly mapped onto the system of annihilating random walkers. This model is quite well understood [44]. Unfortunately, in any dimension larger than one such equivalence is no longer valid.

A slightly more formal description goes as follows: The voter model is a continuous-time Markov process σ_t which takes values in the configuration space $S = \{-1, +1\}^{\Lambda}$. For any $\sigma \in S$ denote as σ^x the state which is obtained from σ by flipping the state of site $x \in \Lambda$, so $\sigma^x(y) = (1 - 2\delta_{xy})\sigma(y)$. It is necessary to know the set of nearest neighbors of x on the lattice Λ. In the case of a d-dimensional hypercubic lattice, there are $2d$ neighbors obtained by shifting the point x by either $+1$ or -1 along the d Cartesian axes. Denote μth neighbor of x by x^{μ}. The transition rates describing a single flip are

$$w(\sigma, \sigma^x) = \frac{1}{2}\left[1 - \sigma(x)\frac{1}{2d}\sum_{\mu=1}^{2d}\sigma(x^{\mu})\right] \tag{3}$$

while all other transition rates are zero

$$w(\sigma, \sigma') = 0\,, \quad \text{if} \quad |\{x \in \Lambda\colon \sigma(x) \neq \sigma'(x)\}| \neq 1\,. \tag{4}$$

The dynamics proceeds according to the master equation

$$\frac{\mathrm{d}}{\mathrm{d}t}p_t(\sigma) = \sum_{x\in\Lambda}\left[w(\sigma^x, \sigma)p_t(\sigma^x) - w(\sigma, \sigma^x)p_t(\sigma)\right]\,. \tag{5}$$

Solution It is possible to directly write the equation for the average state on a single site $S(x, t) \equiv \langle\sigma_t(x)\rangle = \sum_{\sigma\in S} p_t(\sigma)\sigma(x)$, starting with the master equation (5). The result is

$$\frac{\mathrm{d}}{\mathrm{d}t}S(x, t) = \Delta_x S(x, t)\,, \tag{6}$$

where the notation $\Delta_x f(x) = -f(x) + \frac{1}{2d}\sum_{\mu=1}^{2d} f(x^{\mu})$ is used for the discrete Laplace operator on the d-dimensional hypercubic lattice. Equation (6) is in fact the discrete diffusion equation, describing the movement of a random walker over the lattice.

For two-site correlations $R(x - y, t) \equiv \langle\sigma_t(x)\sigma_t(y)\rangle$, the equation is

$$\frac{\mathrm{d}}{\mathrm{d}t}R(x, t) = 2\Delta_x R(x, t) \tag{7}$$

for $x \neq 0$. If the two points coincide, it holds trivially $R(0, t) = 1$ for all times $t \geq 0$. Technically, this is the boundary condition for the solution of the discrete diffusion Eq. (7). In this way it is possible to get closed equations for correlations of all orders.

Equation (6) can be solved by standard mathematical techniques, including Fourier and Laplace transforms. The initial condition is as follows: at the origin the state is $+1$, while all other sites are $+1$ or $+1$ with equal probability, so $S(x, 0) = \delta_{x0}$. The solution is

$$S(x, t) = \mathrm{e}^{-t}\prod_{\alpha=1}^{d} I_{x_\alpha}\left(\frac{t}{d}\right)\,, \tag{8}$$

where $I_{\nu}(z)$ is the modified Bessel function [45].

Asymptotic behavior of the Bessel function gives that, for large times, the average state of any site decays to zero as $S(x, t) \sim t^{-1/2}$, $t \to \infty$. Recall that the initial condition was $S(x, 0) = \delta_{x0}$. So, the average state of the site at the origin decays to zero monotonically, while the other sites, $x \neq 0$, exhibit first an increase in $S(x, t)$ and then decay at later times. This can be understood as propagation of a diffusive wave of $+1$'s from the origin to the rest of the lattice, eventually vanishing at large times.

More information on the dynamics is contained in the two-site correlation function $R(x, t)$. Apart from the factor 2, it obeys the same equation as $S(x, t)$. However, there is an important difference. Equation (7) holds only for $x \neq 0$ and, besides the initial condition $R(x, 0) = \delta_{0x}$, the solution must also satisfy the boundary condition $R(0, t) = 1$. Nevertheless, it is possible, again, to proceed by Fourier- and Laplace-transforming the Eq. (7). The boundary condition enters through the yet unknown function

$$n_{+-}(t) = \frac{1}{2}\left(1 - \frac{1}{2d}\sum_{\mu=1}^{2d} R(x - x^{\mu}, t)\right)\,, \tag{9}$$

which is simply the concentration of interfaces, i. e., the fraction of the bonds connecting sites with unequal state, also called *active bonds*. This quantity is an important measure of the level of activity in the system, as changes of the configuration can occur only at sites adjacent to active bonds.

Let us now separately discuss the results in dimensions $d = 1$, $d = 2$, and $d \geq 3$. In one dimension, the density of interfaces can be expressed in closed form

$$n_{+-}(t) = \frac{1}{2}\mathrm{e}^{-2t}[I_0(2t) + I_1(2t)] \,. \qquad (10)$$

For large times, the behavior is $n_{+-}(t) \simeq (4\pi t)^{-1/2}$ and $\lim_{t\to\infty} R(x,t) = 1$. This means that the activity measured by $n_{+-}(t)$ slowly decays to zero and eventually all the agents become fully correlated. Both features are signatures of complete ordering, which is the fate of the voter model in one dimension.

For $d = 2$ the situation is similar. Again, the density of interfaces decays to zero and the correlation function approaches 1 for large times. However, the evolution is much slower. The solution can be expressed in closed form using the so-called elliptic integrals, but let us mention here only the asymptotic behavior

$$n_{+-}(t) \simeq \frac{\pi}{2}\frac{1}{\ln(16t)} \,, \quad t \to \infty \,. \qquad (11)$$

The behavior changes dramatically for $d \geq 3$, because $n_{+-}(t)$ has a positive limit for $t \to \infty$. In short, the activity never ceases and the voter model never reaches a totally ordered state. Explicit calculations show that the stationary density of interfaces becomes

$$n_{+-}(\infty) = \left(2d\int_0^\infty \left[\mathrm{e}^{-t}I_0(t)\right]^d \mathrm{d}t\right)^{-1} \,. \qquad (12)$$

The asymptotic behavior of the Bessel function guarantees that the integral converges for $d > 2$ and $n_{+-}(\infty) > 0$. Numerical values for some values of d are listed in Table 1 below. One can also find how the density of interfaces approaches its asymptotic value, $n_{+-}(t) - n_{+-}(\infty) \sim t^{1-(d/2)}$.

All the above results for the voter model tacitly assume an infinitely large lattice. In practice, e. g., in numerical simulations, the system always consists of a finite number, $N = L^d$, of sites. This implies that eventually the dynamics lead to one of the two absorbing states, even for $d \geq 3$, where the *infinite* system never orders. The time needed to reach any of the absorbing states in a particular realization of the process is called *stopping time* τ_{st}. The typical scale for the stopping time, called *consensus time* $\tau(N)$, diverges in the thermodynamics limit. It is possible to show rigorously [46] that the consensus time grows like $\tau(N) \sim N^2$ in $d = 1$, $\sim N \ln N$ in $d = 2$ and $\sim N$ for $d \geq 3$. These formulae can be understood on an intuitive level, at least in one dimension, where the dynamics is equivalent to diffusion of domain walls, annihilating upon encounter. The size of homogeneous areas grows, therefore, as the mean displacement of a random walk, $\sim t^{-1/2}$. The time needed to cover the whole system of size N is therefore $\sim N^2$. A similar consideration is also feasible in higher dimensions [36].

The asymptotics of the voter model on d-dimensional hypercubic lattice are summarized in Table 1. Where the behavior is indicated by "$\sim$", the meaning is $t \to \infty$ or $x \to \infty$ or $N \to \infty$, according to the context.

Local Majority Models

Galam Model In democracy, consensus on an issue is rarely achieved by simply waiting until one of the options pervades the whole system through pairwise contact of individuals. Instead, there are various hierarchical levels of decision making, each of which comes to a conclusion based on the principle of majority. This is supposed to lead to a state in which most people are satisfied, when the opinion of the majority of those who participate in the decision process is declared a law. If there were only a single hierarchical level, namely a public referendum (Switzerland may serve as a model example), one could be quite sure that the outcome really represents the majority opinion in the society. However, as soon as there are more levels, and decisions on lower level are passed on to levels above, there is no obvious guarantee that the opinions are not distorted or even reversed. Serge Galam devised a simple model demonstrating that the distortions may not be an exception, but rather a rule [47,48,49,50,51].

The reason for breaking the society into several hierarchical levels is that the cost of communication in a too large collective is prohibitive. It is interesting that this is also the main technical obstacle hindering the introduction of secure electronic elections. One can imagine voting for a presidential candidate from one's home computer over the Internet, but to make the protocol reliable enough to ensure proper secrecy *and* resistance to any attempt at fraud on the scale of an entire nation seems to be beyond currently available technical capabilities [52].

So, the model takes it for granted that the society is organized hierarchically and, on each level, the decision is made within a small group, say, of 3 people. Within each group, majority rule tells what opinion shall be held by the representative of the group when sent to make decision on the upper level. Now, suppose there are only two possible choices, A and B and the fraction of people with opinion A on the level l is $n_A(l)$. On the basic level, $n_A(0)$ represents the concentration of A in the whole population.

Supposing that the people are not correlated in any way, the concentrations $n_A(l)$ provide full information on the system. Essentially, this is a kind of mean-field approx-

Social Processes, Physical Models of, Table 1

Table of some properties of the voter model on a hypercubic lattice in *d* spatial dimensions

d	$n_{+-}(\infty)$	$n_{+-}(t)-n_{+-}(\infty)$	$R(x,\infty)$	$\tau(N)$		d	$n_{+-}(\infty)$
1	0	$\sim t^{-1/2}$	1	$\sim N^2$		3	0.329731...
2	0	$\sim (\ln t)^{-1}$	1	$\sim N \ln N$		4	0.403399...
≥ 3	> 0	$\sim t^{1-d/2}$	$\sim \lvert x\rvert^{2-d}$	$\sim N$		5	0.432410...
						∞	1/2

imation, neglecting any social structure or network within one hierarchical level. The model of majority decisions on a complete graph, also studied in the literature [53,54], is to large extent equivalent to the Galam model.

The dynamics of the model consists in determining the fraction of A at level $n+1$ on the basis of the concentration on level n. For groups of size 3 it leads to the recurrence relation

$$n_A(l+1) = n_A^3(l) + 3\,(1-n_A(l))\,n_A^2(l)\,. \qquad (13)$$

The evolution according to this rule is depicted in Fig. 7a. There are three fixed points at values $n_{A\,\text{fix}} = 0, 1$, and $1/2$, which can be easily checked by insertion in (13). The two extreme values are stable, while the middle point is an unstable fixed point. This reminds us of physical systems with a phase transition, treated by renormalization group techniques [55]. In that formalism, the unstable fixed point corresponds to the critical value of a control parameter and the stable fixed points represent the various possible phases. Here there are two phases, populated uniformly by either all A or all B opinions, constituting total consensus on the issue.

The transition occurs at the symmetric point $n_{A\,\text{fix}} = 1/2$, where exactly half of the people have opinion A. In this sense, the decision making is "fair" because it leads to consensus according to the initial majority. However, a seemingly minor modification makes a big difference. Imagine that the groups formed at each level are composed of an even number of persons, e.g., four. The decision must be taken as to what to do if exactly half of the group pushes for decision A but the rest for B. No majority emerges and the tie must be resolved. If there is an arbitrarily small bias towards one of the choices, say, B, the recurrence relation for the concentrations of opinion A is

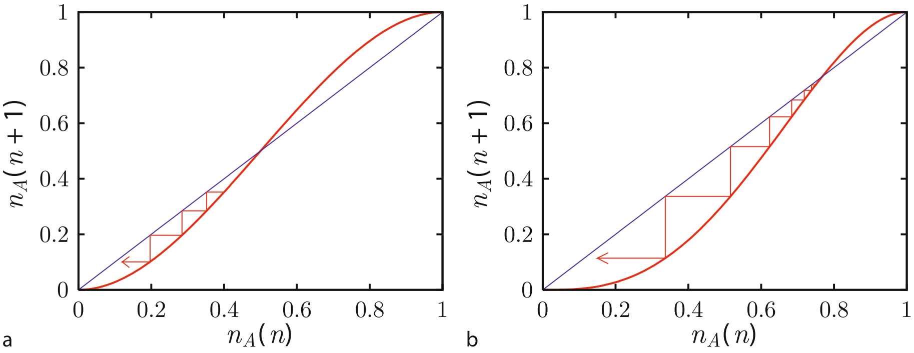
$$n_A(l+1) = n_A^4(l) + 4\,(1-n_A(l))\,n_A^3(l)\,. \qquad (14)$$

Figure 7b shows what happens. The unstable fixed point is shifted significantly to higher values of n_A, namely to

$$n_{A\,\text{fix}} = \frac{1+\sqrt{13}}{6} = 0.76759\ldots \qquad (15)$$

which means that opinion B may win even if it was initially in a minority!

The situation can be even more complicated if some of the people systematically vote against the majority, instead of following the crowd. Such agents are called *contrarians* [56,57]. For simplicity, take (again) groups of size 3. First, the majority option within the group in question is



Social Processes, Physical Models of, Figure 7

Galam model. In a, graph of the recurrence relation for groups of size 3. The *stairs* indicate how the iterations of the relation drive the system away from the unstable fixed point $n_{A\,\text{fix}} = 1/2$. In b, an analogous scheme for groups of size 4 with infinitesimal bias in favor of option *B*. Note the shift of the unstable fixed point to value $n_{A\,\text{fix}} = (1+\sqrt{13})/6$

found, then each member of the group flips her opinion with probability p_c, which is to be interpreted as the concentration of contrarians in the population. The recursion relation is modified to

$$n_A(l+1) = (1-p_c)\left[n_A^3(l) + 3\,(1-n_A(l))\,n_A^2(l)\right] + p_c\left[(1-n_A(l))^3 + 3(1-n_A(l))^2 n_A(l)\right] \tag{16}$$

and results in a shift of the stable fixed points away from the endpoints of the interval $[0,1]$. This means that total consensus is never reached; the contrarians always introduce a certain level of dissidence, but generally the final decision respects the initial majority option in the society. For $p_c < 1/6$ this picture remains valid, as there are three fixed points, with the unstable one keeping its position at $n_{A\,\mathrm{fix}} = 1/2$. However, if the concentration of contrarians rises and $p_c > 1/6$, the three fixed points coalesce to a single stable fixed point at $n_{A\,\mathrm{fix}} = 1/2$. No consensus is ever reached and the distribution of opinions tend to a precisely equilibrated state of equally represented opinions A and B. Several recent cases of popular referenda or presidential elections ending in extremely narrow victories come to mind, showing that the presence of contrarians may have palpable consequences for our lives.

A natural question is, how fast is consensus approached as one climbs higher and higher up ladder of hierarchies? It is evident that as we start closer to the unstable fixed point, it takes us more time to approach one of the stable points. To get a quantitative estimate, let us turn to the simplest case described by Eq. (13) and replace the discrete level index l by a continuous variable l which can be interpreted as time elapsed during the formation of the consensus. Thus, the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}l} n_A(l) = -n_A(l)(n_A(l)-1)(2n_A(l)-1) \tag{17}$$

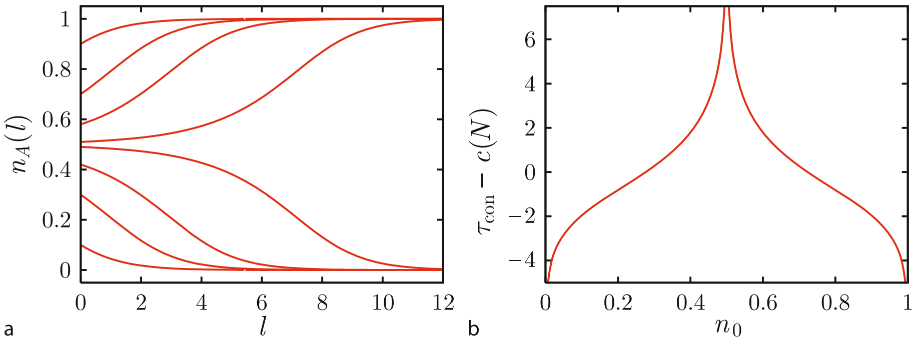
is obtained, which can be solved relatively easily. The solution for several initial conditions is shown in Fig. 8a. Of course, starting from any point $n_A(0)$ inside the interval $(0,1)$, the time to reach either of the stable fixed points 0 or 1 is infinite. However, in reality it is not necessary to reach the fixed point exactly, because the population consists of a finite number N of people, so full consensus is reached if the procedure is stopped at a distance $1/N$ from the fixed point. The time to achieve that diverges for increasing N as $\sim \ln N$, as will be seen from the explicit calculation.

To be more precise, the stopping time t_{st} as a function of the initial concentration $n_0 \equiv n_A(0) < 1/2$ can be defined by the formulae $n_A(0) = n_0$ and $n_A(t_{\mathrm{st}}(n_0)) = 1/N$, expressing the initial and final conditions, respectively. The case $n_0 > 1/2$ differs only in the final condition, which is $n_A(t_{\mathrm{st}}(n_0)) = 1 - 1/N$. Knowing the general solution of (17) the formula is

$$t_{\mathrm{st}}(n_0) = \ln\frac{n_0(1-n_0)}{(2n_0-1)^2} + \ln\frac{(N-2)^2}{N-1}\,. \tag{18}$$

Hence the already-mentioned logarithmic divergence of the stopping time for $N \to \infty$. The N-independent part of the stopping time can be seen in Fig. 8b. Note the divergence for $n_0 \to 1/2$, which can be regarded as a sign of a certain kind of dynamic phase transition.

The Galam model has potential for extensions in various directions and, indeed, it was further generalized [58,59,60,61,62,63,64,65,66,67,68]. For us, the most physically relevant question is how the dynamics change



Social Processes, Physical Models of, Figure 8

Galam model with continuous levels, for group size 3. The evolution according to Eq. (17) is shown in a, with initial concentration of opinion A equal $n_0 = 0.9, 0.7, 0.58, 0.51, 0.49, 0.42, 0.3$, and 0.1 (from *top* to *bottom*). In b, the time to reach consensus in a society of N people is shown. In this graph, $c(N) = \ln\frac{(N-2)^2}{N-1}$

if the people are distributed on a fixed lattice or network [69,70,71]. After all, the Galam model has all the attributes of a mean-field version of some more complicated model, even though it may not be obvious which one [72]. The next section will make the point more clear.

Sznajd Model Clearly, a group of people puts stronger pressure on an individual than each member of the group taken separately. Nature designed humans so that they follow the crowd. If you see two or more people sharing an opinion on certain issue, you are tempted to join. This is the basic idea behind the model invented by Katarzyna Sznajd-Weron and Józef Sznajd (the daughter and the father) [73]. In its first version, the model was defined on a one-dimensional lattice of length N. Each site is inhabited by an agent which can be in two states, denoted $+1$ and -1, as in the voter or majority-rule models. It may correspond to people choosing between two dominant brands of a certain product in the market or voting in a two-party political system. In each step of the dynamics, a pair of neighbors is chosen randomly. If they are in the same state, say, $+1$, then the two sites adjacent to the pair adopt the same opinion $+1$, propagating the consensus outwards. Conversely, if they differ in opinion, they propagate the dissensus. The rule is shown schematically in Fig. 9.

Formally, suppose the chosen pair is in state $(\sigma(x), \sigma(x+1))$, at time t. If $\sigma(x) = \sigma(x+1)$, the neighbors of the pair are updated as $\sigma_{t+1/N}(x-1) = \sigma_{t+1/N}(x+2) = \sigma(x)$, while for $\sigma(x) \neq \sigma(x+1)$ the update rule is $\sigma_{t+1/N}(x-1) = \sigma(x+1)$, $\sigma_{t+1/N}(x+2) = \sigma(x)$. Unlike the voter or majority rule model, there are three absorbing states. Besides the obvious "ferromagnetic" states of all $+1$ or all -1, there is the "antiferromagnetic" state where the sites in states $+1$ and -1 alternate regularly. Strictly speaking, there are two such states, one characterized by $+1$ at sites with odd coordinates and the other by $+1$ at even sites.

Looking at the rules more thoroughly, it is evident that the linear chain can be divided into two sublattices, one of them containing all odd coordinates x, the other all even sites. The state of the agents in one sublattice is never influenced by the agents in the other one. Moreover, the dynamics within one sublattice are a trivial modification of the voter model. The only difference is that in the voter model, one site induces change of state of one of its neighbors, while here the agent induces *both* of its neighbors adopt its state. Indeed, the update rule can be written as $\sigma_{t+1/N}(x-1) = \sigma(x+1)$, $\sigma_{t+1/N}(x+2) = \sigma(x)$, irrespective of the relation between $\sigma(x)$ and $\sigma(x+1)$ [74]. This observation simplifies the model to a large extent. Indeed, it is equivalent to two copies of the voter model evolving in parallel, where the only coupling between them comes from the fact that the sites to be updated are neighbors on the original lattice. The two voter models evolve so that the update occurs at the same place in both copies. However, the initial state of both sublattices is chosen randomly and independently and many properties, for example the concentrations of $+1$ in either of the sublattices, remain uncorrelated forever.

Restricting the study to one sublattice only, the solution follows the same way as was used successfully for the voter model. The flipping rates are analogous to (3), differing only in factor 2, originating from the fact that a site induces a change in state of its two neighbors. This factor means only a rescaling of time, so it is quite safe to neglect it. The result for the voter model can be translated directly to our case.

As a first application, let us look at the asymptotic states. The probability of reaching the state of all $+1$ is simply the concentration of $+1$ opinions in the initial state. This holds for both sublattices independently. From here, it is straightforward to deduce the probabilities P_+, P_- and P_{AF} that the system ends in absorbing states with all $+1$, all -1, and the antiferromagnetic state, respectively, as a function of the initial concentration of $+1$ opinions. Indeed, a state with all $+1$ means that both sublattices reached the uniform $+1$ state (and similarly for -1) while the antiferromagnetic state is obtained if one of the sublattices ended in a $+1$ state and the other in a -1 state. Hence

$$\begin{aligned} P_+(n_+) &= n_+^2 \\ P_-(n_+) &= (1-n_+)^2 \\ P_{AF}(n_+) &= 2(1-n_+)n_+ \,. \end{aligned} \tag{19}$$

The history of a single agent is also interesting. During its evolution, it may change its state several (including zero) times. The time elapsed between two subsequent changes of state is called *decision time* t_{dec}. It is natural to ask, what is the probability distribution of decision times of all agents? The answer relies on mapping the evolution of the voter model onto diffusive motion of domain walls. As explained in Sect. "Voter Model", the domain walls separating the regions of $+1$ and -1 opinions, evolve in time like annihilating random walkers. Selecting an agent at site x, its state flips if and only if the domain wall crosses the point x. The decision time is nothing other than the time between two successive visits of the random walker to the same position x, or, using language familiar to experts in random walks, the decision time is the time of first return of the random walk to the origin. Calculation of this quantity can be found in probability theory textbooks.

$$\text{Model of Ref. [73]}\ \left\{\begin{array}{lcl} \left.\begin{array}{lcl}(\cdot\,++\,\cdot) & \longrightarrow & (++++)\\ (\cdot\,--\,\cdot) & \longrightarrow & (----)\end{array}\right\}\ \text{Sznajd model} \\ (\cdot\,-+\,\cdot) \quad \longrightarrow \quad (-+-+) \\ (\cdot\,+-\,\cdot) \quad \longrightarrow \quad (-+-+) \end{array}\right.$$

Social Processes, Physical Models of, Figure 9
Illustration of the dynamics of opinions in the the model of Sznajd–Weron and Sznajd, as it appears in the first paper [73] and as it was modified subsequently and named the "Sznajd model". The dots replace any state of the site

The decision time distribution behaves asymptotically as a power law

$$P^{>}_{\mathrm{dec}}(t) \equiv \mathrm{Prob}\{t_{\mathrm{dec}} > t\} \sim t^{-1/2}\,, \quad t \to \infty \qquad (20)$$

and this is exactly what was found numerically in the founding work of the Sznajds.

The solution obtained by mapping to the voter model is only the beginning of the story. Things start to be exciting again as soon as the "simplification" of the dynamic rules introduced above is made, allowing only neighbor pairs with equal states to influence their neighborhood. If the two agents in the pair do not agree in their opinions, nothing happens. In the scheme shown in Fig. 9 this corresponds to taking only first two rows as allowed updates. It is this modification that has been widely studied, and common consensus has assigned it the name Sznajd model (although calling it "Sznajds' " would perhaps do more justice to the authors).

As a first step in the analysis of the Sznajd model, it is important to note that there are again three absorbing states, all $+1$, all -1 and the antiferromagnetic one. But now the antiferromagnetic state is unstable, because randomly flipping a single site results in a nucleus of three sites in the same state, which irresistibly invades the whole system. But although unstable, the existence of such an absorbing state leaves important traces in the dynamics, as will be more clear later.

The transition rates for the underlying Markov process are

$$\begin{aligned} w(\sigma, \sigma^x) = \frac{1}{8}\Big[&\sigma(x-2)\sigma(x-1) + \sigma(x+1)\sigma(x+2) \\ &- \sigma(x)\big(\sigma(x-2) + \sigma(x-1) + \sigma(x+1) \\ &+ \sigma(x+2)\big) + 2\Big]\,. \end{aligned} \qquad (21)$$

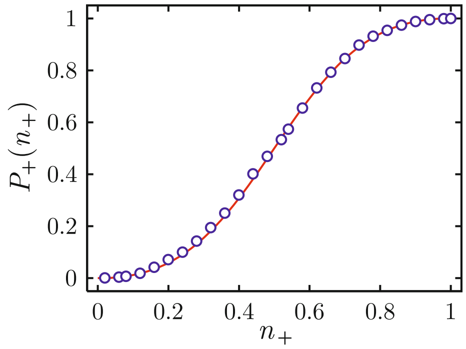
It is instructive to compare this with Eq. (3) describing voter model dynamics. The Sznajd model is a member of the family of non-linear voter models. This means that the transition rates depend non-linearly (quadratically, in fact) on the states of sites other than x.

There is a rather standard technique of approximate solution of similar systems, called the Kirkwood approximation. In our case, its use yields the probability for ending in the all $+1$ state in the form

$$P_+(n_+) = \frac{n_+^2}{(1-n_+)^2 + n_+^2}\,. \qquad (22)$$

Figure 10 compares this result with numerical simulations, showing very good agreement.

Two more questions have been asked regarding the one-dimensional Sznajd model. The first is the decision time, introduced a short while ago. It was found numerically [74] that it follows the same power law (20) as in the previous case, where it resulted from underlying linear voter dynamics. For the Sznajd model, no proof is available showing that this is the exact result, but it is possible to understand this result on an intuitive level. Although the transition rates (21) correspond to a non-linear, rather than linear, voter model, the evolution of domain walls, responsible for the behavior of the decision time, remains very similar. If the domain walls are far from each other, they again perform a random walk. Non-linearity comes into play only when the domain walls come



Social Processes, Physical Models of, Figure 10
One-dimensional Sznajd model. Probability of reaching the final configuration of all sites in state $+1$, depending on the initial concentration n_+ of $+1$ sites. The *points* are numerical simulation data extracted from [75], the *line* is the analytical result (22) obtained using the Kirkwood approximation

close together, more precisely when they are one lattice spacing apart. Then, they can no longer move freely, they can make only one step towards one another, thus annihilating each other. The dynamics is again those of a diffusion-annihilation process, but the domain walls are not independent; in addition to annihilation they interact at a short distance.

The interaction is in fact a trace of the unstable antiferromagnetic absorbing state. Two walls at distance one form a small nucleus of antiferromagnetic state, thus slowing down the dynamic with respect to free annihilating random walkers. If by chance several domain walls come so close that the distance between neighboring walls is 1, they form an antiferromagnetically ordered cluster, where the walls squeezed between another two walls from both the left and right sides cannot move at all and the dynamics is hindered even more.

However, for large times, the density of domain walls is low and the influence of such antiferromagnetic islands can be neglected. Therefore, it is no big surprise that the power-law dependence of the decision time remains unaffected, at least for large times.

The second interesting question is how many agents never changed their opinion up to certain time t, or, what is the *persistence* in the dynamics of the Sznajd model? In the voter model in one dimension, there is an exact analytical solution [44] showing that the fraction of sites which retain their initial opinion at least to time t decays as $\sim t^{-3/8}$. It may seem quite surprising that the same behavior was found numerically in the one-dimensional Sznajd model [76,77]. However, applying the same arguments as above, showing that the long-time dynamics is essentially dominated by annihilating random walks of the domain walls, it follows that it is quite plausible that the long-time behavior of the persistence is governed by the same exponent 3/8 both in the voter and the Sznajd models in one dimension.

Agents organized on a line are now understood, and it is natural to ask what changes if they are arranged in a two-dimensional mesh, like the people scattered on the surface of the Earth. For example, on a square lattice, one-dimensional dynamics can be generalized in such a way that a pair of neighboring agents is chosen, and if the two have the same state, they make all their 6 nearest neighbors share their state [78,79]. Slightly more formally, let x be a random coordinate on a square lattice with $N = L^2$ sites and $y = x^\nu$ one of the 4 neighbors of x. If $\sigma_t(x) = \sigma_t(y)$, then the configuration is updated as $\sigma_{t+1/N}(x^\mu) = \sigma_{t+1/N}(y^\mu) = \sigma_t(x)$ for all $\mu \in \{1,2,3,4\}$. Numerical simulations show an interesting difference from the one-dimensional case. The probability $P_+(n_+)$ has discontinuity at $n_+ = 1/2$, indicating a dynamical phase transition, while in one dimension the dependence is smooth. More strongly, the data suggest that $P_+(n_+) = 0$ for $n_+ < 1/2$ and $P_+(n_+) > 0$ in the opposite case $n_+ > 1/2$. This result is equally valid in any dimension larger than 1. In the limit of infinite dimension, the behavior should coincide with the mean-field approximation, which implies putting the Sznajd model on a complete graph [54]. Such a modification is analytically solvable, as we now show.

On a complete graph with N vertices, the configuration of the system is fully described by the number N_+ of agents in state $+1$ or by the magnetization $m = 2N_+/N - 1$. The evolution of the latter is governed by transition rates which in the limit $N \to \infty$ yield the Fokker–Planck equation

$$\frac{\partial}{\partial t} P_m(m,t) = -\frac{\partial}{\partial m}\left[\frac{1}{2}(1-m^2)mP_m(m,t)\right] \tag{23}$$

for the probability density $P_m(m,t)$ of the magnetization at time t. Before proceeding further, note that the average magnetization $\langle m\rangle(t) = \int mP_m(m,t)\mathrm{d}m$ satisfies the equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle m\rangle(t) = \frac{1}{2}\left(\langle m\rangle(t) - \langle m^3\rangle(t)\right)\ , \tag{24}$$

which has a structure similar to Eq. (17) describing the dynamics of the Galam model. Indeed, equating $m = 2n_A - 1$, identifying the level index with time, $l = t$, and making the approximation $\langle m^3\rangle = \langle m\rangle^3$, thus neglecting fluctuations, the Eqs. (17) and (24) coincide. So, the Galam and Sznajd models are solved in one shot.

It can be easily verified that the general solution of Eq. (23) has the form

$$P_m(m,t) = \frac{1}{(1-m^2)m} f\left(\mathrm{e}^{-t/2}\frac{m}{\sqrt{1-m^2}}\right) \tag{25}$$

for arbitrary function $f(y)$. The latter has to be determined from initial conditions, and assuming that at the beginning the magnetization was m_0 for all realizations of the process, it occurs that $f(m/\sqrt{1-m^2}) = (1-m^2)m\delta(m-m_0)$. Clearly, in this case the probability density consists of a single δ-function which moves, as time passes, towards one of the ends of the allowed interval for the magnetization, $m \in [-1,1]$. The dynamics is those of a pure deterministic drift and no diffusion ever smears out the evolving probability packet, as can be seen immediately from the absence of a second-derivative term in Eq. (23). Clearly, this is due to the limit of infinite system size and for finite N there is an additional diffusive term in the Fokker–Planck equation, proportional to N^{-1}.

The deterministic nature of the evolution of opinions has profound consequences. First, the initial sharp δ-function distribution of magnetization remains sharp until the end, thus justifying the neglect of fluctuations and the replacement $\langle m^3 \rangle = \langle m \rangle^3$. The fluctuations are set on only at times which diverge when $N \to \infty$ and the equivalence to the Galam model is exact in the thermodynamic limit. In a finite system of size N, the typical time t_{st} to reach the absorbing state, starting from magnetization m_0, can be estimated by requiring that drift brings the magnetization to the distance of order $\sim 1/N$ from either of the extremal points $m = \pm 1$. Expressed in terms of the initial concentration of the $+1$ opinion, $n_+ = (m_0 + 1)/2$, it is

$$t_{\mathrm{st}} \simeq \ln \left(\frac{(1 - n_+) n_+ N}{(2 n_+ - 1)^2} \right) . \qquad (26)$$

So, having clarified the close relation to the Galam model, it comes as no surprise that the Eqs. (26) and (18) coincide for large N.

For any positive initial magnetization the final state is always the uniform configuration of all agents in state $+1$, and vice versa. The probability of ending in the all $+1$ state is the step function $P_+ = \theta(n_+ - 1/2)$. This is consistent with simulations of the Sznajd model in dimensions larger than 1 and confirms the existence of a dynamical phase transition at $n_+ = 1/2$. It also shows that in a society where everybody interacts with everybody, but no person changes her opinion unless she meets at least two other people who compel her to do so, the initial majority, however narrow it is, always takes all. No chance is left to minorities. Fortunately enough, reality is more complex. The next section describes model implementations of one common "complication", that the number of choices is not limited to two but can be large, or even very large.

Bounded Confidence

Axelrod Model

Definition One of the features characteristic of the way culture is shared and propagated around the globe is that similar cultures are much more prone to mutual convergence, while incompatible lifestyles often coexist side by side without visibly influencing each other. Robert Axelrod introduced a model [80] nicely describing such a situation.

In the Axelrod model, contrary to the voter or Sznajd model, the character of each of the agents is given by more than one feature. One can think of tastes regarding food, sports, music, etc. These categories represent the features. For each feature the taste can assume various values, e. g., somebody likes eating raw vegetables, spending whole days in the fitness center and listening to Mozart in the evenings, while somebody else feeds on French fries, watches football on TV and adores the pop-star of the season. If two neighbors do not agree on any feature, they are so different that they do not influence each other. Conversely, if they find at least one feature where they share the same preference, one of them may find a second feature in which they do differ and change the preference on that second feature so that it agrees with the preference of the neighboring agent. The fact that the agents do not always interact, but do so only if they have something in common, is called *bounded confidence*.

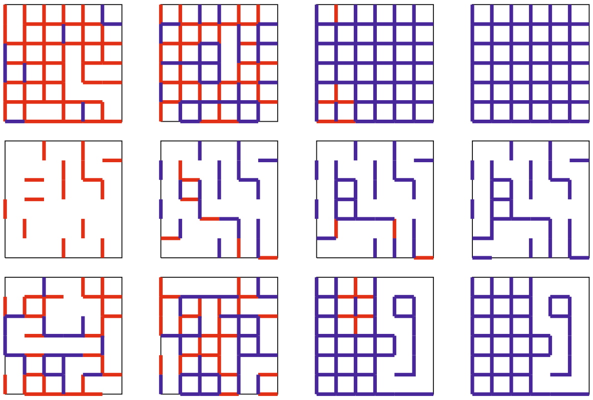
More precisely, there are $N = L^d$ agents placed again on the d-dimensional hypercubic lattice $\Lambda = \{0, 1, \ldots, L-1\}^d$ with periodic boundary conditions and endowed with $F > 1$ integers. The numbers represent the values of its F features, so the state of the agent at site $x \in \Lambda$ is described by the vector with coordinates $\sigma(x, i) \in \mathbb{Z}$, $i = 1, 2, \ldots, F$. The configuration space of the model is $\mathbb{Z}^{\Lambda \times \{1,\ldots,F\}}$ and the evolution of the configuration σ_t is a Markov process determined by transition rates

$$w(\sigma, \sigma^{x,f,a}) = \frac{1}{2d} \sum_{\mu=1}^{2d} \theta \left(A(x, x^\mu) \right) \times \left[F - A(x, x^\mu) \right]^{-1} \delta_{a, \sigma(x^\mu, f)} \, , \qquad (27)$$

where the summation goes over the set of $2d$ neighbors x^μ of the site x. We denoted here $\sigma^{x,f,a}$ the configuration which differs from σ only in feature f, so $\sigma^{x,f,a}(x, f) = a$ and $\sigma(x, f) \neq a$. The function $A(x, y) = \sum_{g=1}^{F} \delta_{\sigma(x,g), \sigma(y,g)}$ counts the number of features on which the agents at positions x and y agree, and $\theta(x)$ is the Heaviside function, $\theta(x) = 1$ for $x > 0$ and zero otherwise. The first factor after the sum in (27) accounts for the condition that the neighbors must agree in at least one feature and the second factor is here due to the fact that in each update the agent chooses randomly among $F - A(\sigma(x), \sigma(x^\mu))$ features in which she disagrees with her neighbor.

Simulations If the two factors after the sum in (27) were absent, the transition rates would depend linearly on the states of the neighbors and a generalization of the linear voter model would emerge, with all its beautiful solubility. However, the non-linearity is there and makes the model non-trivial. Lacking suitable analytical tools, it is necessary to rely on numerical simulations.

To be precise, the rules of the dynamics of the Axelrod model implemented in simulations are as follows: On



Social Processes, Physical Models of, Figure 11
Typical evolution of the Axelrod model on a square lattice with periodic boundary conditions. In each row, snapshots of the configuration in four times $t_1 < t_2 < t_3 < t_4$, from *left* to *right*, where the last configuration is the absorbing state. The active bonds are drawn in *red*, the bonds with full consensus in *blue*, the bonds with absolute disagreement are left *white*. The initial condition is drawn from $\mathbb{Z}_q$ with uniform probability. The parameters are, from *top* to *bottom*, $F = 3$ and $q = 2$; $F = 3$ and $q = 14$; $F = 2$ and $q = 2$. The frame in *light gray* around each configuration is there only for visual convenience

a lattice of size $N = L^d$, time t is discrete and proceeds in chunks of $1/N$, so from time t to $t+1$ there are as many elementary updates as there are sites. One update step consists of:

1. Choose a site x at random; randomly choose one if its neighbors x^μ. Count the number A of features in which the agent at x agrees with the chosen neighbor.
2. If $A = 0$, do nothing. If $0 < A < F$, randomly choose a feature f among those in which the agent and the neighbor differ. Then set the value of the feature f equal, $\sigma_{t+1/N}(x, f) = \sigma_t(x^\mu, f)$.

One of the principal quantities of interest will be the density of active bonds, which is a generalization of the density of interfaces investigated in the voter model. A bond connecting agents at sites x and y is called active if the agents differ in at least one feature and also agree in at least one feature, i. e., $0 < A(\sigma(x), \sigma(y)) < F$. Their fraction relative to the total number of bonds in the lattice will be denoted $n_A(t)$.

When the number of active bonds drops to zero, after the stopping time τ_{st}, evolution freezes and the system reaches one if its absorbing states. It means that all neighbors are either identical or absolutely incompatible. Clusters of agents which share the same values of all features can be identified, while the borders of such clusters are marked by bonds with no shared value. Even for the simplest case of two features $F = 2$ and the set of allowed values for these features constrained to only two elements, there are infinitely many possible absorbing states characterized by various cluster configurations. (For a finite lattice this number is of course finite, but grows very fast with the system size.) This leads to an important note: Speaking of culture, there might be a good many stable configurations and it is impossible (and improper) to discriminate as to which is the best one.

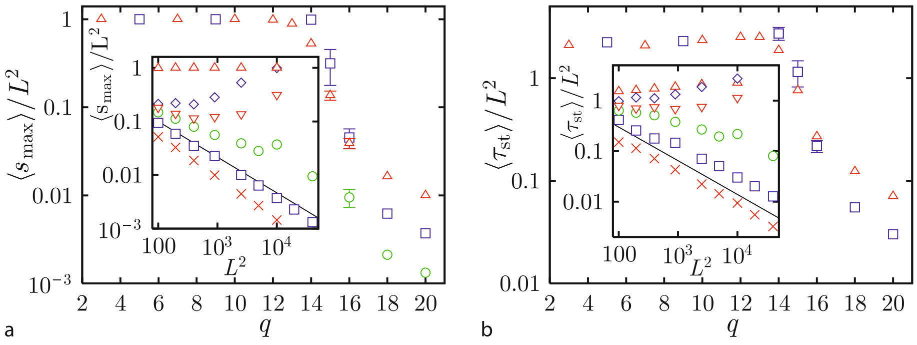
A full characterization of the set of absorbing states would be too difficult. A rough measure is the number of agents in the largest cluster $s_{\max}$, compared to the total number of agents N. When this figure is close to the total system size, it means that the absorbing state is uniform, or very close to it. If the size of the largest cluster is small, or even remains finite when $N \to \infty$, the absorbing state is "multicultural" or fragmented into many isolated islands without mutual interaction, much like the hundreds of villages in Papua-New Guinea, each of them having its own particular language.

The evolution and shape of the eventual absorbing state depends critically on the initial condition. First, note that if, at the beginning, all the values of the features belonged to certain subset $V \subset \mathbb{Z}$ of integers, all features will remain within the set V forever. The most important thing is how large the set V is. But even when V is infinite, the values can be mostly concentrated on a finite subset of V, very rarely going beyond it. A simple but reliable measure of the effective size of the set V, taking into account how often each member is actually present, is the *inverse participation ratio* (IPR). Suppose that, in the initial state, the probability of some feature of a randomly picked agent having the ith value from the set V is p_i. Then, the IPR is

$$q^{-1} = \sum_{i=1}^{|V|} p_i^2 \,. \tag{28}$$

Such a definition is motivated by a simple consideration: if set V consists of the q lowest non-negative integers, i. e., $V = \mathbb{Z}_q \equiv \{0, 1, \ldots, q-1\}$, and all members have the same probability $p_i = 1/q$, the inverse participation ratio is exactly q^{-1}. The integer q is simply the number of elements which are participating, hence the name.

Phase Transition Let us look at how the Axelrod model behaves on a two-dimensional square lattice. Recall that the linear size of the lattice is L, so there are $N = L^2$ agents. The initial condition will consist in choosing the values of all features for all agents independently from the same uniform distribution on $\mathbb{Z}_q$. The typical evolution is shown in



Social Processes, Physical Models of, Figure 12

Phase transition in Axelrod model on a square lattice of size $L \times L$, for $F = 3$, when changing the localization parameter q of the initial state. In a, relative size of the largest cluster, in b, average stopping time relative to the system size are shown. Different symbols correspond to sizes $L = 50$ (△), $L = 100$ (□), and $L = 200$ (○). In the *insets*, dependence of the corresponding quantities on the system size L^2 are shown. Symbols correspond to parameters $q = 5$ (△), 14 (◇), 15 (▽), 16 (○), 20 (□), and 30 (×). The *straight line* is the power $\sim L^{-4/3}$

Fig. 11. For larger F and q the absorbing state is very fragmented, but decreasing q it becomes totally uniform and a single cluster covers the entire lattice. On the other hand, for small F the absorbing state contains a few moderately-sized clusters.

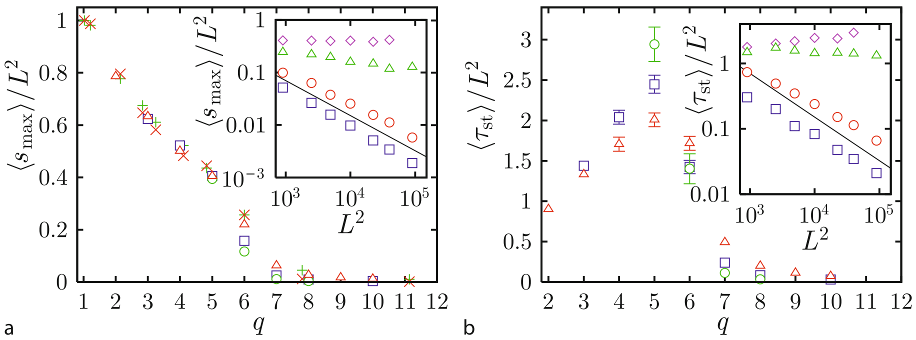
These vague observations are put on a quantitative basis in Figs. 12 and 13. Let us look first at the behavior of the average size of the largest cluster, $\langle s_{\max}\rangle$, when the parameter q is changed. The most important finding is that, at a certain value $q = q_c$, there is a phase transition separating the regime $q < q_c$, in which the maximum cluster makes up a finite fraction of the whole lattice, from the phase with $q > q_c$, where the fraction of sites within the largest cluster goes to zero for $N \to \infty$. The quantity $\langle s_{\max}\rangle/N$ can serve as an order parameter.

Another quantity which illustrates phase transition is the average stopping time, i. e., the time to reach the absorbing state. Generically, it grows with the system size as $\langle \tau_{st}\rangle \sim N^{\eta}(q)$, but below the transition it is proportional to the system size, so $\eta(q) = 1$, while for $q > q_c$ it grows more slowly, with the exponent $\eta(q) \simeq 1/3$. Interestingly, comparison of the stopping time with the size of the largest cluster leads to the conclusion that, for $F = 3$, the two quantities are roughly proportional, $\langle \tau_{st}\rangle \propto \langle s_{\max}\rangle$, both of them exhibiting a jump at the transition, suggesting that the clusters grow linearly in time until they reach the absorbing state. On the other hand, for $F = 2$, such proportionality is not observed. Instead, the average maximum cluster decreases continuously to zero when q_c is approached from below and at the same time the stopping time seems to diverge.

With discrete q, it is impossible to determine the position of the phase transition with high precision, but in practice it is possible to make some reasonable estimates, based on the dependence of $\langle s_{\max}\rangle/N$ on N, as shown in the insets. For $F = 3$ the critical value of the parameter q lies somewhere close to $q_c \simeq 15$, while for $F = 2$ the estimated value is $q_c \simeq 5$.

A more important difference between cases $F = 2$ and $F = 3$ than the numerical value of the transition point q_c is the very nature of the phase transition. It was already shown that the stopping time normalized to the system size blows up at the transition only for $F = 2$, remaining finite for $F = 3$. With fair trustworthiness it is possible to say that for $F = 2$ the transition is continuous, i. e., second-order, while for $F = 3$ the transition is clearly a first-order one. Although it may be dangerous to rely only on numerical data for the order parameter, the conclusion on the nature of the phase transition was also supported by results for the distribution of cluster sizes close to the transition. In Fig. 14a it can be seen that the probability of finding in the absorbing state a cluster of size larger than or equal to s decays as a power, i. e., $P^{>}_{\text{clust}}(s) \sim s^{1-\alpha}$, with an obvious cutoff at large s due to the finiteness of the lattice. The value of the exponent α plays critical role. For $F = 2$ the estimate is $\alpha_{F=2} \simeq 1.6$ and for $F = 3$ it is $\alpha_{F=3} \simeq 2.65$. It has been shown that the exponent does not depend on the number of features as long as $F \geq 3$ [81]. How does the distribution of cluster sizes relate to the type of the phase transition? Denoting $N_{\text{clust}}(N)$ the total number of clusters and $P_{\text{clust}}(s, N) = P^{>}_{\text{clust}}(s) - P^{>}_{\text{clust}}(s + 1)$ the fraction of clusters of size s on the lattice with N sites, the result is

$$N = N_{\text{clust}}(N) \sum_{s=1}^{N} s P_{\text{clust}}(s, N) \,. \tag{29}$$



Social Processes, Physical Models of, Figure 13

Phase transition in the Axelrod model on a square lattice of size $L \times L$, for $F = 2$, when changing the localization parameter q of the initial state. In a, relative size of the largest cluster, in b, average stopping time relative to the system size. In both panels, results for the uniform initial distribution is plotted for sizes $L = 50$ (△), 100 (□), 200 (○). In the *insets*, dependence of the corresponding quantities on the system size, for $q = 5$ (◇), 6 (△), 7 (○), and 8 (□). The *straight lines* are power laws $\sim L^{-4/3}$ In a, there are also data extracted from [81], where the initial condition was assembled from Poisson-distributed integers; sizes were $L = 50$ (+) and $L = 150$ (×)

If $P_{\rm clust}(s, N) \sim s^{-\alpha}$ and $\alpha < 2$, the sum on the right hand side diverges when $N \to \infty$. But the sum is simply the average cluster size, the quantity which plays the role of correlation length. The behavior resembles the percolation transition and the diverging correlation length is an unmistakable characteristic of a second-order phase transition.

On the other hand, if $P_{\rm clust}(s, N) \sim s^{-\alpha}$ with $\alpha > 2$, the sum converges. To keep the equality in (29) for $N \to \infty$, either the number of clusters must be proportional to the system size, or, in addition to the power-law complement, there must be an additional term in the distribution of cluster sizes, accounting for the largest cluster of size $s_{\max} \simeq N$, i. e.,

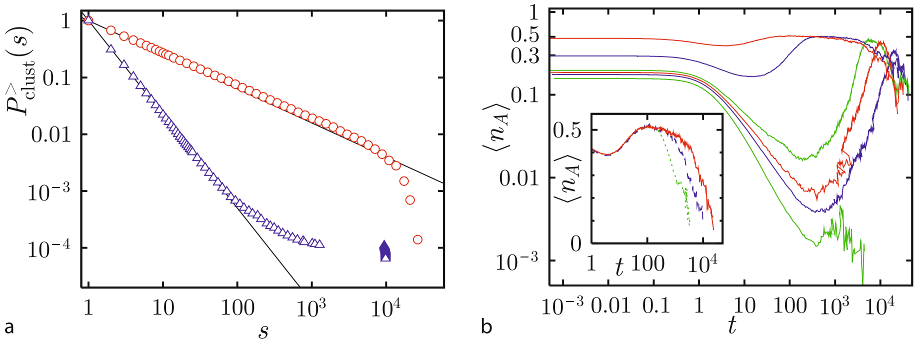
$$P_{\rm clust}(s, N) \simeq a s^{-\alpha} + b\ \ \delta_{s,s_{\max}}\ . \tag{30}$$

The former possibility occurs for $q > q_{\rm c}$ and the latter for $q < q_{\rm c}$. Indeed, the simulation data in Fig. 14a show the presence of a δ-function part in (30), with positive weight $b > 0$. Below the transition, the largest cluster spans essentially whole system, and at $q_{\rm c}$ its size drops discontinuously to a value negligible with respect to N. This is typical of first-order transitions.

These considerations are also compatible with the behavior of the stopping time as a function of q. For $F = 2$, data indicate that $\langle \tau_{\rm st}\rangle/N$ diverges as $q \to q_{\rm c}$, while for $F = 3$ it remains constant. Interpreting the stopping time as the correlation time of the dynamics, the model exhibits exactly the same behaviors commonly observed in the second- and first-order phase transitions, respectively.

To conclude, the phase transition in the Axelrod model is driven by the parameter q, measuring the localization of values in the initial condition, and belongs to the class of first-order transitions, if the number of features is at least 3, while for only 2 features the transition is continuous. Interestingly, the same dependence in terms of number of components is known in the Potts model. However, one should bear in mind that the transition in the Axelrod model has a purely dynamical origin. There is no equilibrium besides the absorbing states, so the analogy with phase transitions in equilibrium statistical physics must be taken cautiously and with certain reserve.

Although the most interesting question touches the properties of the absorbing states, it is instructive, also, to see how the system approaches them. In Fig. 14b the evolution of the average number of active bonds is shown. The averages are taken only over realizations which have not yet reached the absorbing state. Around the time $t \simeq 1$, i. e., after as many updates on the computer as there are lattice sites, the density of interfaces decreases drastically. When $q > q_{\rm c}$ it brings the system quickly to the absorbing state, while in the opposite case, $q < q_{\rm c}$, the activity rises again and n_A increases to a value close to $n_A \simeq 0.5$. Then, a very slow evolution follows. Most often, the evolution ends due to an occasional hit into an absorbing state, but this is, rather, a finite size effect, as demonstrated in the inset in Fig. 14b. When system size increases, slow evolution is increasingly prolonged. To see the behavior at these very long times is very difficult on a computer, but it seems that the density of active bonds decays very slowly to zero, resembling the logarithmic decay of the density of interfaces in the two-dimensional voter model.



Social Processes, Physical Models of, Figure 14
Axelrod model on a square lattice. In a, distribution of cluster sizes close to the transition $q \simeq q_c$, averaged over 100 realizations. The parameters are $F = 2$, $q = 5$, $L = 200$ (◯), and $F = 3$, $q = 15$, $L = 100$ (△). The *straight lines* are power laws $\sim s^{-0.6}$ and $\sim s^{-1.65}$. In the data for $F = 3$, note the the isolated points around $s \simeq 10^4 = L^2$ indicating that the distribution is composed of a power-law part plus a δ-function located close to L^2. In b, time dependence of the number of active bonds, averaged over realizations of the process. The lattice size is $L = 100$, number of features $F = 3$ and different lines correspond to $q = 5, 9, 14, 15, 16$, and 18, from *top* to *bottom*. In the *inset*, the dynamics is shown for $F = 3$, $q = 5$, and three lattice sizes, $L = 30$ (*dotted line*), 50 (*dashed line*), and 100 (*solid line*)

Bounded Confidence with Continuous Opinions

There may be a different view on consensus formation, assuming that opinions on a certain issue can vary continuously, but a discussion on that subject and possible convergence of the opinions cannot take place unless the actual opinions are close enough. If the people differ too much, they may even avoid mutual contact completely or decide to use "different means", an often-used euphemism for killing the opponent. If, on the contrary, the difference in their opinions, measured by a continuous variable, does not exceed a certain threshold, then the difference may be further diminished by discussions and peaceful persuasion. This idea is called *bounded confidence* and a kind of it is also a key ingredient of the Axelrod model.

Linear Dynamics Modeling consensus formation using continuous variables has been around for quite a long time. Let us denote F_i as the variable describing the opinion of the individual i. The simplest way to implement the convergence of opinions, introduced by DeGroot [82], is to assume that in an upcoming period, opinions are linear combinations of the current opinions of all individuals

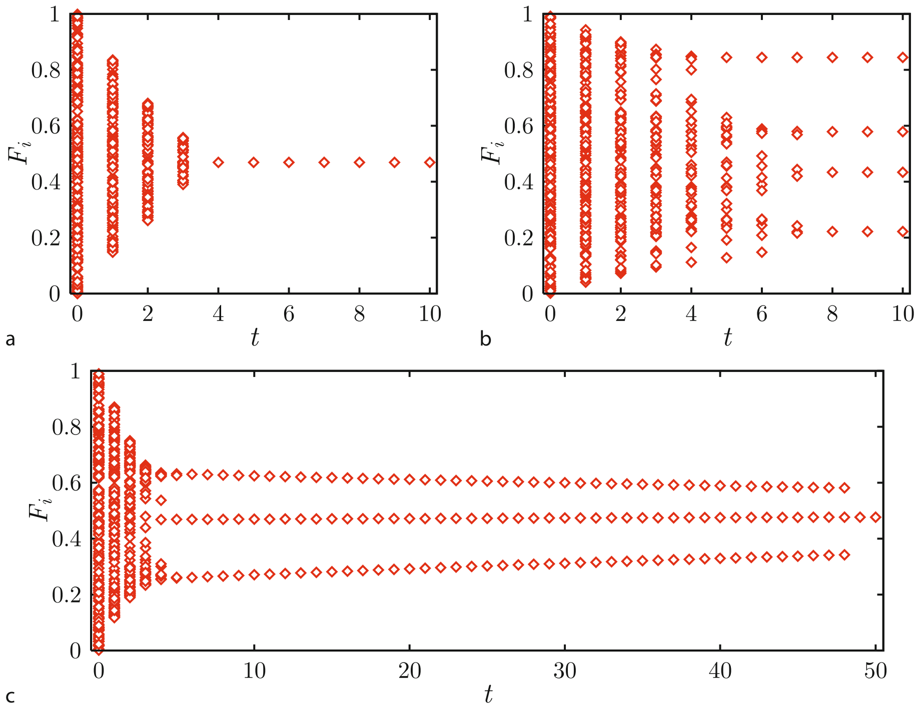
$$F_i(t+1) = \sum A_{ij} F_j(t) \,, \tag{31}$$

where A is a stochastic matrix, i. e., all its elements are non-negative, $A_{ij} \geq 0$, and the sum of all its columns is one, $\sum_i A_{ij} = 1$. This is called a *confidence matrix*. The diagonal elements are assumed to be strictly positive, $A_{ii} > 0$, which means that the individuals have at least some non-vanishing belief in their own opinions. The new values of F_i are weighted averages of previous opinions, the weights being stored in the matrix A. As such, the dynamics is very simple and is equivalent to those of a Markov chain with transition probabilities A_{ij}.

In DeGroot's model, the question of reaching a consensus or not is reduced to the study of connected components of the graph of direct interactions. When there are more of them around, each of them reaches its own consensus, independent of the others. This is quite simple. Introducing bounded confidence adds complications.

Hegselmann and Krause The new opinion of individual i takes into account others' opinions with weights A_{ij}. What if these weights are not constant, but depend on the current opinions themselves? Although some attempts in this direction have been made earlier [83], the principle of bounded confidence was first applied in the models developed by the groups of Deffuant et al. [84] and Krause and Hegselman [85,86]. Let us discuss the latter model first.

The Hegselmann–Krause (HK) model [86] considers opinion formation as a fully deterministic process, as in DeGroot's model investigated in the previous section. The new opinion of an individual is the average of current opinions of those (including the individual of concern) whose opinions lie within a fixed confidence bound $\epsilon > 0$ from the individual's own opinion.



Social Processes, Physical Models of, Figure 15
Evolution of the Hegselmann–Krause model. Each point represents one or more agents with specified value F_i. Total number of agents is $N = 200$. In a, the confidence bound is $\epsilon = 0.3$, while in b it is $\epsilon = 0.1$. c shows how the approach to consensus is slowed down when the confidence bound is close to its critical value. Here $\epsilon = 0.24$

For those pairs which do not differ more than ε in their opinions, the weights A_{ij} are uniform, so

$$A_{ij} = \begin{cases} 0 & \text{for} \quad |F_i - F_j| > \epsilon \\ \frac{1}{N_{i\epsilon}} & \text{for} \quad |F_i - F_j| \le \epsilon \,. \end{cases} \tag{32}$$

Obviously, the normalization constant is the number of individuals within the confidence bound, $N_{i\epsilon} = |\{j\colon |F_i - F_j| \le \epsilon\}|$. Note that the dependence of the confidence matrix on current opinions is very strongly non-linear. From the formal point of view, the most important thing is that the zero-pattern of the confidence matrix can also change in time. It is possible to interpret it as changing the structure of the graph of directly communicating individuals. To see if consensus is going to be reached, it is important to observe whether the graph remains connected during evolution. Clearly, once it splits into disconnected parts, it will never join together again and consensus will not be achieved.

To get an impression of how the evolution of the HK model proceeds, look at Fig. 15. For a large enough confidence bound ε, the system approaches full consensus, while lower values of ε induce several stable communicating subgroups which do not interact with each other, and the system splits into several clusters with different opinions. The number of such clusters grows as $1/\epsilon$ when the confidence bound shrinks. Numerical simulations show that the critical value ϵ_c, at which the full consensus breaks down, approaches the value about $\epsilon_c \simeq 0.2$ when the number of individuals increases. Interestingly, when the confidence bound comes close to the critical value ϵ_c, the number of steps needed to reach consensus increases, suggesting that this is indeed a kind of a dynamical phase transition. This behavior is rather robust, as it persists in various modifications of the HK model, most notably if the individuals can communicate only through a regular lattice or a random network [87,88,89,90].

Deffuant et al. Apart from the initial condition, the model of Hegselmann and Krause is fully deterministic and opinions are updated in parallel. A similar model, but with random sequential update, was introduced by G. Deffuant et al. [84].

There are, again, N individuals with opinions F_i. In each update step, two of them, say, i and j, are chosen randomly. Then, we check to see whether their opinions differ less than (or equally to) the confidence bound ε. In the positive case, their opinions are slightly shifted towards each other

$$\left.\begin{aligned} &F_i(t+1/N)\\ &\quad=(1-\mu)F_i(t)+\mu F_j(t)\\ &F_j(t+1/N)\\ &\quad=\mu F_i(t)+(1-\mu)F_j(t) \end{aligned}\right\} \text{for} \quad |F_i(t)-F_j(t)| \le \epsilon \tag{33}$$

where μ is a parameter fixing the rate of convergence.

For very large numbers of individuals, $N \to \infty$, the dynamics can be expressed in terms of the continuous distribution of opinions $P(f,t)$, which can be written formally as $\int_0^f P(f',t)\mathrm{d}f' = \lim_{N\to\infty} \frac{1}{N}\sum_j \theta(f-F_j(t))$. The following innocent-looking rate equation is obtained [91]

$$\begin{aligned} \frac{\partial}{\partial t}P(f,t) = &\int\limits_{|f_1-f_2|\le\epsilon} P(f_1,t)P(f_2,t)\\ &\times\left[\delta((1-\mu)f_1+\mu f_2-f)-\delta(f_1-f)\right]\mathrm{d}f_1\mathrm{d}f_2\,, \end{aligned} \tag{34}$$

which exhibits fairly complex behavior. Starting from the uniform initial condition $P(f,0)=1$ for $f\in[0,1]$, the configuration evolves into a stationary state composed of one or several δ-functions, each of them corresponding to one cluster of individuals sharing the same opinion.

To see this in a simple example, consider the case $\epsilon > 1$. All individuals interact with each other, so their opinions converge to a common limit $f = 1/2$. It is possible to see this when we use Eq. (34) to investigate the time evolution of the moments of the distribution. Obviously $\frac{\mathrm{d}}{\mathrm{d}t}\langle f\rangle = 0$, so the average opinion is independent of time, $\langle f\rangle = 1/2$. The dispersion from the mean $\langle f^2\rangle_\mathrm{c} = \langle f^2\rangle - \langle f\rangle^2$ obeys

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle f^2\rangle_\mathrm{c} = -2\mu(1-\mu)\langle f^2\rangle_\mathrm{c}\,, \tag{35}$$

which means that the dispersion decays exponentially to zero with rate $\mu(1-\mu)$. This confirms our intuition, noted earlier, that μ determines the speed of the evolution towards the stationary state.

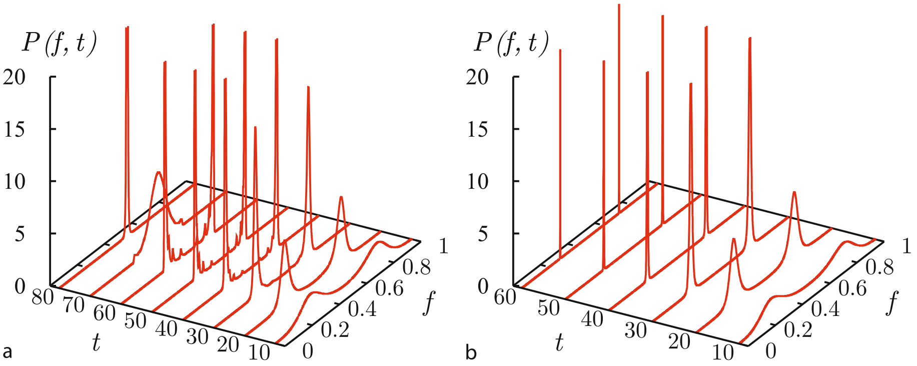
The behavior in the complementary regime of very small ε can be guessed, assuming that $P(f,t)$ does not vary too wildly at scales comparable to ε or shorter. This also means that f is assumed to be farther than ε from the extremal values 0 and 1. In this case, $P(f_{1,2},t)$ is expanded in a Taylor series and the integral on the right hand side of (34) can be performed explicitly. Finally, the following partial differential equation is obtained

$$\frac{\partial}{\partial t}P(f,t) = -\frac{\epsilon^3}{3}\mu(1-\mu)\frac{\partial^2}{\partial x^2}P^2(f,t)\,. \tag{36}$$

There is a trivial homogeneous solution $P(f,t) = C$, independent of f and t, which is, however, unstable at all length scales. To see it, it is enough to linearize (36) close to the uniform solution, $P(f,t) = C + D(f,t)$, where $|D(f,t)| \ll C$, and express the result in terms of the Fourier transform $\widetilde{D}(k,t)$ of the small perturbation. The result is $\frac{\partial}{\partial t}\widetilde{D}(k,t) = \frac{2}{3}k^2\epsilon^3 C\mu(1-\mu)\widetilde{D}(k,t)$, indicating that perturbations increase as their wavelength decreases. On the other hand, if a stationary state is reached, the structures created cannot be closer to each other than ε. This feature is lost in the derivation of Eq. (36), but nevertheless the conclusion is that, for small ε, the stationary state will be composed of regularly spaced δ-functions with period proportional to ε.

To see explicitly how the stationary state is approached for general values of the confidence bound ε, the solution of Eq. (34) should be found numerically. Examples of such evolution are shown in Fig. 16. For large enough ε, all opinions converge to the common value 1/2, as in the case $\epsilon > 1$. However, low ε produces two or more distinct peaks. Detailed analysis [91,92] reveals that a series of bifurcations occurs in the system when the confidence bound is decreased. The consensus breaks into separated groups suddenly at a critical value ϵ_{c1}, which break themselves into more groups at $\epsilon_{c2} < \epsilon_{c1}$ and so forth. The numerically found value of the first critical point is $\epsilon_{c1} = 0.269\ldots$ But how do the peaks emerge from the originally homogeneous distribution? It can be seen in Fig. 16 that they are triggered by the original inhomogeneity at the edges of the interval [0, 1]. A structure of gradually sharpening peaks propagates towards the middle of the interval, where non-linear waves from opposite endpoints meet and form the resulting pattern.

This observation can be translated into common language by saying that the ultimate fate of the opinions in the society is largely in hands of the extremists. Indeed, individuals close to the edges of the scale of possible opinions are, from the very beginning, drawn towards the middle of the interval [0, 1], i. e., close to the centrist position, while those who are farther from the edges than ε keep their opinions unchanged for some finite time. The crucial question is, how far from the extreme will the former extremists move before they settle? It may happen that they will drift so close to the midpoint that the centrists will attract them all; full consensus follows. But the extremists themselves can attract enough centrists and, instead of creating one peak at the center, two (or even more) peaks oc-



Social Processes, Physical Models of, Figure 16
Evolution of the Deffuant model with $\mu = 0.5$. In a, approach to consensus for confidence bound $\epsilon = 0.27$, slightly larger than the critical value ϵ_{c1}. In a, two peaks are formed in the distribution of opinions, for $\epsilon = 0.25$

cur. The resulting society lives with polarized (or diversified) opinions. The parameter which decides the outcome is the confidence bound ε. Higher confidence leads to an overwhelming consensus among people, while lower confidence causes the society split into several non-communicating communities.

Emergence of Social Classes

People are not equal. Not only do they differ in the color of their eyes, in their ability to play chess or to run a marathon race, but even individuals with very similar talents may find themselves very different in social status. All attempts to bring more justice into such evident disequilibrium has ended in desperate or even catastrophic failure. Perhaps the best one can do is to make the membranes separating social levels as permeable as possible, so that no one is a priori disqualified. The ubiquity of social stratification in animal as well as human collectives is certainly a phenomenon which calls for an explanation, and the fact that rabbits, dogs, apes and *Homo sapiens* exhibit similar behavior suggests some common mechanisms which may not be too complicated after all, although they produce highly complex outcomes.

Bonabeau Model

A newcomer in an animal group always has to undergo some fighting before its placement in the social ladder is commonly accepted. If, on the other hand, an individual leaves the group for some prolonged period and returns back, it has to fight again, as the previously established level has faded away. These two observations motivated Bonabeau et al. [93] to introduce a model of self-organized hierarchies [93,94,95,96,97,98].

To be clear, the hierarchies are not understood in the sense of trees with a king or a marshal on the top and lesser ranks below. Instead, it means an ordering, each individual bearing a single number called strength, indicating its position among others. When two agents meet, the stronger one has higher probability to be strengthened, while the weaker is most likely pushed down even more. The strengths of the agents who do not meet at this time relax towards zero by a fixed fraction. More formally, the configuration of the system at time t is described by the collection of strengths $F_i(t)$ of the agents $i = 1, 2, \ldots, N$. In each step, a pair of agents is chosen to fight, i and j, say, resulting in a change of the strengths by ± 1 for the winner and loser, respectively. Moreover, all strengths relax to the reference (zero) level deterministically. Thus

$$\begin{aligned} F_i\left(t+\frac{1}{N}\right) &= \left(1-\frac{\mu}{N}\right) F_i(t) + \Delta_{ij} \\ F_j\left(t+\frac{1}{N}\right) &= \left(1-\frac{\mu}{N}\right) F_j(t) - \Delta_{ij} \\ F_k\left(t+\frac{1}{N}\right) &= \left(1-\frac{\mu}{N}\right) F_k(t)\,, \quad k \neq i\,,\ k \neq j\,, \end{aligned} \tag{37}$$

with

$$\text{Prob}\left\{\Delta_{ij} = \pm 1\right\} = \frac{1}{1 + \mathrm{e}^{\mp\eta(F_i(t)-F_j(t))}}\,. \tag{38}$$

The parameter η tunes the level of randomness in the dynamics, where $\eta \to \infty$ corresponds to purely deterministic outcomes, the stronger agent always beating the weaker, while $\eta = 0$ means that the strengths are increased and decreased by mere chance.

So far, the question of which pairs of agents interact and when, has not been dealt with. The most natural choice is to place the individuals on a network, or simply on a square lattice, leaving some sites empty, and to allow the agents to diffuse along the edges of the network. When two agents happen to meet at one site, they fight. More than two agents on a site is not allowed; they behave like hard hemispheres, each site being able to accommodate one full sphere. Numerical simulations of such systems [93,94,95,96,97] show that the change in the density of agents, or, equivalently, the relaxation rate μ, induces a phase transition from a uniform state with all agents' strength close to 0 to a hierarchical state where the strengths are highly non-homogeneous. The order parameter is the dispersion of the number of fights won by the agents, i. e.,

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(w_i - \frac{1}{2}\right)^2}\,,$$

where $w_i = n_i^+/(n_i^+ + n_i^-)$ and n_i^+, n_i^- are the number of encounters won and lost by the agent i, respectively. In the homogeneous phase $\sigma = 0$ while in the hierarchical phase it has a finite value.

Mean-Field Solution

Analytical study is possible in the mean-field approximation. Indeed, if diffusion is fast enough to ensure many encounters with various agents during the typical time given by the speed of relaxation of the strengths towards zero, the spatial structure of the lattice on which the diffusion takes place becomes irrelevant. In other words, at each time step, two agents are chosen at random and allowed to fight. In this case the dynamics is much simpler, and for a large system the following deterministic evolution equations for the strengths is obtained:

$$\frac{\mathrm{d}F_i}{\mathrm{d}t} = -\mu F_i + \frac{1}{N}\sum_{j=1}^{N}\frac{\sinh\eta(F_i - F_j)}{1 + \cosh\eta(F_i - F_j)}\,. \tag{39}$$

The stochastic term decreases as $1/N$ for large N, so it is neglected here.

It can be easily seen that the average strength $\overline{F} = \frac{1}{N}\sum_i F_i$ relaxes exponentially to 0 according to $\frac{\mathrm{d}\overline{F}}{\mathrm{d}t} = -\mu\overline{F}$. So, it is sufficient to consider only stationary states with zero mean, $\overline{F} = 0$. The simplest of them, satisfying (39), is the trivial uniform state $F_i = 0$. This may be unstable, though, and its linear stability must be investigated.

Using the notation $\frac{\mathrm{d}F_i}{\mathrm{d}t} = R_i(F_1, \ldots, F_N)$ for the expression occurring in (39), the eigenvalues of the matrix $H_{ij} = \frac{\partial}{\partial F_j}R_i(0, \ldots, 0)$ should be inspected. Thus

$$H_{ij} = \left(-\mu + \frac{\eta}{2}\right)\delta_{ij} - \frac{\eta}{2N}\,. \tag{40}$$

One of the eigenvectors is uniform, $x_i = 1$, corresponding to eigenvalue $-\mu$, which is always negative. The remaining $N-1$ eigenvectors have the form $x_i = 1 - N\delta_{ik}$ for some k and they all belong to the same eigenvalue $\frac{\eta}{2} - \mu$. If the latter is positive, the uniform solution of the Eq. (39) is unstable. This happens for

$$\mu \le \mu_c \equiv \frac{\eta}{2} \tag{41}$$

and the question naturally occurs, what are the stationary configurations beyond the critical value μ_c? Not very far beyond the critical point, linear stability analysis can offer a useful hint, but it needs to be complemented by the influence of the lowest non-linear terms. Therefore, expanding the right-hand side of (39) up to the third order in the differences $F_i - F_j$, one gets for the stationary state

$$(2\mu - \eta)F_i = \frac{\eta^3}{12N}\sum_j (F_i - F_j)^3\,. \tag{42}$$

The solution is assumed in the form

$$F_i = a\delta_{ik} - b \tag{43}$$

suggested by the eigenvectors of H_{ij} corresponding to unstable modes. Inserting the trial solution (43) into (42) leads to a set of equations for the parameters a and b which can be solved easily

$$a = \begin{cases} 0 & \text{for} \quad \mu > \frac{\eta}{2} \\ \pm\sqrt{\frac{12N(\eta-2\mu)}{(N-2)\eta^3}} & \text{for} \quad \mu < \frac{\eta}{2} \end{cases} \tag{44}$$

$$(2\mu - \eta)b = -\frac{1}{12N}\eta^3 a^3\,. \tag{45}$$

Hence, the order parameter σ is deduced. Note that in the stationary state, the fraction w_i of fights won by the agent i should be balanced by a relaxation of the strength. Indeed, $2w_i - 1$ is the average increase of the strength of agent i in one step, which should be equal to μF_i. So, $\sigma = \mu\sqrt{\sum_i F_i^2/N/2}$ and inserting the result (45) yields

$$\sigma = \begin{cases} 0 & \text{for} \quad \mu > \frac{\eta}{2} \\ \frac{\mu}{\eta}\sqrt{\frac{3}{N}\left(1 - \frac{2\mu}{\eta}\right)} & \text{for} \quad \mu < \frac{\eta}{2}\,, \\ & \mu \to \frac{\eta}{2} \quad \text{and} \quad N \gg 1\,. \end{cases} \tag{46}$$

How can this solution be interpreted? One may compare it to a situation in a society with one master and many servants, a single emperor while all the rest are servants, equal to each other in their subordination. This looks nice, but it follows from the formulae (45) that this "realistic" solution always comes together with a mirror image of itself, a society of a single servant subject to many equal masters. This is an absurd situation, revealing something artificial in the Bonabeau model itself. Indeed, the dynamics is invariant with respect to inversion of all strengths, $F_i \to -F_i$, because no a priori advantage of being stronger was explicitly introduced. The only thing which was assumed was that the strong grow stronger and the weak grow weaker.

Now let us turn back to mathematical aspects of the model. It can be shown that a solution of type (43) exists for any value of μ and η, with only the complication that the equations for a and b become transcendental, thus analytically insoluble. This would be too simple, though. In fact, it is easy to find a stable solution in the limit $\eta \to \infty$, with μ kept finite, and see that it differs completely from (43). The expression within the sum in (39) becomes the sign function of the difference $F_i - F_j$ and the set of equations for the stationary state becomes

$$\mu F_i = \frac{1}{N} \operatorname{sign}(F_i - F_j) \tag{47}$$

and reordering the agents so that their strengths make an increasing sequence $F_1 < F_2 < \ldots < F_N$ one gets

$$F_i = \frac{1}{\mu N}(2i - N - 1)\,, \tag{48}$$

a society organized as a regular ladder of ranks. The order parameter corresponding to such a state,

$$\sigma = \frac{1}{3}\left(1 - \frac{1}{N^2}\right) \tag{49}$$

approaches a non-zero limit for $N \to \infty$, contrary to the configuration of (43) which is appropriate only in close proximity to the critical point, and whose order parameter decreases as $N^{-1/2}$, as can be seen from (44).

Despite the obvious criticism that the essential mirror symmetry $F_i \to -F_i$ is unrealistic and leads to unacceptable solutions, the Bonabeau model manifests quite well the basic idea of how the various social classes emerge. Within this framework, hierarchies are due to dynamical instabilities of the uniform state. The same general mechanism is also (at least partially) responsible for the stratification of human society. Any departure from the mean, be it positive or negative, is amplified. Any accidental misfortune sends one almost invariably even deeper. That is the whole mystery.

Future Directions

The prospects of physical modeling of social behavior are larger than what is presented here. The study of the emergence of cooperation within the framework of spatial and repeated prisoner's dilemma game is one of the classics. It may be, and has been, generalized in various ways. For example, one may wonder what happens if the agents could move through the lattice, or along the edges of the general graph representing the structure of the society. Indeed, people do not sit on the same lattice point forever. It turns out that, depending on the average velocity of the agents' movement, there is a phase transition. Less mobile agents form a phase with a macroscopic fraction of collaborators, while beyond a certain level of mobility the collaborators vanish.

Bounded confidence models pose some subtle and still unsolved problems. One of them concerns how the final state depends on the number of agents, or, stated differently, on the granularity of the distribution of opinions on the opinion axis. If the number of agents were arbitrarily large, the density could be arbitrarily small and it could happen that between any two agents there might be other agents which may mediate ultimate consensus. If that were true, the splitting of society into disparate groups would be the effect of limited size. The conclusion sounds somewhat paradoxical: small societies are more likely to be divided into non-communicating sub-cultures, while large societies turn to be more homogeneous. Another question naturally follows: how small must a social group be to allow disparate opinions to coexist? Implications for multicultural studies are evident.

The dynamics of social strata are relatively less explored and the state of the problem presented here must be considered as preliminary. The societies sketched in these models do not possess any internal structure, which would either stabilize or slow-down the emergence of stratification. Surely there is much space for future work here.

Some important areas were not covered at all, for example the whole family of physically based models of the birth and competition of languages [41,99]. Some social phenomena also closely touch economics, but these are covered by other articles in this Encyclopedia.

Acknowledgments

The original results in this article were obtained within the projects AVOZ10100520 and MSM0021620845.

Bibliography

Primary Literature

1. Comte A (1822) Plan des travaux scientifiques nécessaires pour réorganiser la société
2. Comte A (1839) Cours de philosophie positive, tome IV, 46e leçon. Bachelier Paris

3. Chigier NA, Stern EA (eds) (1975) Collective phenomena and the applicatios of physics to other fields of science. Brain Research Publications, Fayetteville
4. Callen E, Shapero D (1974) A theory of social imitation. Phys Today 27(7):23–28
5. Weidlich W (1991) Physics and social science – The approach of synergetics. Phys Rep 204:1–163
6. Anderson PW, Arrow KJ, Pines D (1988) The economy as an evolving complex system. Addison Wesley, Reading
7. Galam S (2004) Sociophysics: A personal testimony. Physica A 336:49–55
8. Galam S, Gefen Y, Shapir Y (1982) Sociophysics: A new approach of sociological collective behaviours. I. Mean-behaviour description of a strike. J Math Sociol 9:1–13
9. Galam S, Moscovici S (1991) Towards a theory of collective phenomena: Consensus and attitude changes in groups. Eur J Soc Psychol 21:49–74
10. von Neumann J, Morgenstern O (1944) Theory of games and economic behavior. Princeton University Press, Princeton
11. Vega-Redondo F (1996) Evolution, games, and economic behaviour. Oxford University Press, Oxford
12. Nash JF (1950) Equilibrium points in n-person games. Proc Natl Acad Sci USA 36:48–49
13. Nash JF (1950) The bargaining problem. Econometrica 18:155–162
14. Marsili M, Zhang YC (1997) Fluctuations around Nash equilibria in game theory. Physica A 245:181–188
15. Axelrod R (1980) Effective choice in the prisoner's dilemma. J Confl Resolut 24:3–25
16. Axelrod R (1980) More effective choice in the prisoner's dilemma. J Confl Resolut 24:379–403
17. Axelrod R, Hamilton WD (1981) The evolution of cooperation. Science 211:1390–1396
18. Lindgren K (1991) Evolutionary phenomena in simple dynamics. In: Langton CG, Taylor C, Farmer JD, Rasmussen S (eds) Artificial Life II. Addison-Wesley, Readwood City, pp 295–312
19. Lindgren K (1997) Evolutionary dynamics in game-theoretic models. In: Arthur WB, Durlauf SN, Lane DA (eds) The economy as an evolving complex system II. Perseus, Reading, pp 337–367
20. Nowak MA, May M (1992) Evolutionary games and spatial chaos. Nature 359:826–829
21. Schweitzer F, Behera L, Muhlenbein H (2002) Evolution of cooperation in a spatial prisoner's dilemma. Adv Compl Syst 5:269–299
22. Szabó G, Töke C (1998) Evolutionary prisoner's dilemma game on a square lattice. Phys Rev E 58:69–73
23. Chiappin JRN, de Oliveira MJ (1999) Emergence of cooperation among interacting individuals. Phys Rev E 59:6419–6421
24. Lim YF, Chen K, Jayaprakash C (2002) Scale-invariant behavior in a spatial game of prisoner's dilemma. Phys Rev E 65:026134
25. Szabó G, Vukov J, Szolnoki A (2005) Phase diagrams for an evolutionary prisoner's dilemma game on two-dimensional lattices. Phys Rev E 72:047107
26. Abramson G, Kuperman M (2001) Social games in a social network. Phys Rev E 63:030901(R)
27. Ebel H, Bornholdt S (2002) Evolutionary games and the emergence of complex networks. arXiv:cond-mat/0211666 (Preprint)
28. Ebel H, Bornholdt S (2002) Coevolutionary games on networks. Phys Rev E 66:056118
29. Zimmermann MG, Eguíluz VM (2005) Cooperation, social networks, and the emergence of leadership in a prisoner's dilemma with adaptive local interactions. Phys Rev E 72:056118
30. Vukov J, Szabó G, Szolnoki A (2006) Cooperation in noisy case: Prisoner's dilemma game on two types of regular random graphs. cond-mat/0603419
31. Clifford P, Sudbury A (1973) A model for spatial conflict. Biometrika 60:581–588
32. Holley RA, Liggett TM (1975) Ergodic theorems for weakly interacting infinite systems and the voter model. Ann Prob 3:643–663
33. Liggett TM (1985) Interacting particle systems. Springer, Berlin
34. Redner S (2001) A guide to first-passage processes. Cambridge University Press, Cambridge
35. Scheucher M, Spohn H (1988) A soluble kinetic model for spinodal decomposition. J Stat Phys 53:279–294
36. Krapivsky PL (1992) Kinetics of monomer-monomer surface catalytic reactions. Phys Rev A 45:1067–1072
37. Frachebourg L, Krapivsky PL (1996) Exact results for kinetics of catalytic reactions. Phys Rev E 53:R3009–R3012
38. Ben-Naim E, Frachebourg L, Krapivsky PL (1996) Coarsening and persistence in the voter model. Phys Rev E 53:3078–3087
39. Dornic I, Chaté H, Chave J, Hinrichsen H (2001) Critical coarsening without surface tension: The universality class of the voter model. Phys Rev Lett 87:045701
40. Al Hammal O, Chaté H, Dornic I, Muñoz MA (2005) Langevin description of critical phenomena with two symmetric absorbing states. Phys Rev Lett 94:230601
41. Castellano C, Fortunato S, Loreto V (2007) Statistical physics of social dynamics. arXiv:0710 3256
42. ben-Avraham D, Considine D, Meakin P, Redner S, Takayasu H (1990) Saturation transition in a monomer-monomer model of heterogeneous catalysis. J Phys A: Math Gen 23:4297–4312
43. Liggett TM (1999) Stochastic interacting systems: Contact, voter, and exclusion processes. Springer, Berlin
44. Derrida B, Hakim V, Pasquier V (1996) Exact exponent for the number of persistent spins in the zero-temperature dynamics of the one-dimensional Potts model. J Stat Phys 85:763–797
45. Gradshteyn IS, Ryzhik IM (1994) Table of integrals, series, and products, 5th edn. Academic Press, San Diego
46. Cox JT (1989) Coalescing random walks and voter model consensus times on the torus in $\mathbb{Z}^d$. Ann Prob 17:1333–1366
47. Galam S (1986) Majority rule, hierarchical structures, and democratic totalitarianism: A statistical approach. J Math Psychol 30:426–434
48. Galam S (1990) Social paradoxes of majority rule voting and renormalization group. J Stat Phys 61:943–951
49. Galam S (1999) Application of statistical physics to politics. Physica A 274:132–139
50. Galam S (2000) Real space renormalization group and totalitarian paradox of majority rule voting. Physica A 285:66–76
51. Galam S, Wonczak S (2000) Dictatorship from majority rule voting. Eur Phys J B 18:183–186
52. Schneier B (1996) Applied cryptography, 2nd edn. Wiley, New York

53. Krapivsky PL, Redner S (2003) Dynamics of majority rule in two-state interacting spin systems. Phys Rev Lett 90:238701
54. Slanina F, Lavička H (2003) Analytical results for the Sznajd model of opinion formation. Eur Phys J B 35:279–288
55. Plischke M, Bergersen B (1994) Equilibrium statistical physics. World Scientific, Singapore
56. Galam S (2004) Contrarian deterministic effects on opinion dynamics: The hung elections scenario. Physica A 333:453–460
57. Stauffer D, Sá Martins JS (2004) Simulation of Galam's contrarian opinions on percolative lattices. Physica A 334:558–565
58. Florian R, Galam S (2000) Optimizing conflicts in the formation of strategic alliances. Eur Phys J B 16:189–194
59. Galam S (2002) Minority opinion spreading in random geometry. Eur Phys J B 25:403–406
60. Galam S (2002) The September 11 attack: A percolation of individual passive support. Eur Phys J B 26:269–272
61. Galam S (2003) Modelling rumors: The no plane Pentagon french hoax case. Physica A 320:571–580
62. Galam S (2003) Global physics: From percolation to terrorism, guerilla warfare and clandestine activities. Physica A 330:139–149
63. Galam S, Mauger A (2003) On reducing terrorism power: A hint from physics. Physica A 323:695–704
64. Galam S, Vignes A (2005) Fashion, novelty and optimality: An application from Physics. Physica A 351:605–619
65. Galam S (2004) The dynamics of minority opinions in democratic debate. Physica A 336:56–62
66. Galam S (2004) Unifying local dynamics in two-state spin systems. cond-mat/0409484
67. Galam S (2005) Local dynamics vs. social mechanisms: A unifying frame. Europhys Lett 70:705–711
68. Gekle S, Peliti L, Galam S (2005) Opinion dynamics in a three-choice system. Eur Phys J B 45:569–575
69. Galam S, Chopard B, Masselot A, Droz M (1998) Competing species dynamics: Qualitative advantage versus geography. Eur Phys J B 4:529–531
70. Tessone CJ, Toral R, Amengual P, Wio HS, San Miguel M (2004) Neighborhood models of minority opinion spreading. Eur Phys J B 39:535–544
71. Galam S (2005) Heterogeneous beliefs, segregation, and extremism in the making of public opinions. Phys Rev E 71:046123
72. Sousa AO, Malarz K, Galam S (2005) Reshuffling spins with short range interactions: When sociophysics produces physical results. Int J Mod Phys C 16:1507–1517
73. Sznajd-Weron K, Sznajd J (2000) Opinion evolution in closed community. Int J Mod Phys C 11:1157–1165
74. Behera L, Schweitzer F (2003) On spatial consensus formation: Is the Sznajd model different from a voter model? cond-mat/0306576
75. Krupa S, Sznajd-Weron K (2005) Relaxation under outflow dynamics with random sequential updating. Int J Mod Phys C 16:177–1783
76. Stauffer D, de Oliveira PMC (2002) Simulation of never changed opinions in Sznajd consensus model using multi-spin coding. cond-mat/0208296
77. Stauffer D, de Oliveira PMC (2002) Persistence of opinion in the Sznajd consensus model: Computer simulation. Eur Phys J B 30:587–592
78. Stauffer D, Sousa AO, Moss de Oliveira S (2000) Generalization to square lattice of Sznajd sociophysics model. Int J Mod Phys C 11:1239–1245
79. Bernardes AT, Costa UMS, Araujo AD, Stauffer D (2001) Damage spreading, coarsening dynamics and distribution of political votes in Sznajd model on square lattice. Int J Mod Phys C 12:159–167
80. Axelrod R (1997) The dissemination of culture: A model with local convergence and global polarization. J Confl Resolut 41:203–226
81. Castellano C, Marsili M, Vespignani A (2000) Nonequilibrium phase transition in a model for social influence. Phys Rev Lett 85:3536–3539
82. DeGroot MH (1974) Reaching a consensus. J Am Stat Assoc 69:118–121
83. Chatterjee S, Seneta E (1977) Toward consensus: Some convergence theorems on repeated averaging. J Appl Prob 14:89–97
84. Deffuant G, Neau D, Amblard F, Weisbuch G (2000) Mixing beliefs among interacting agents. Adv Compl Syst 3:87–98
85. Krause U (2000) A discrete nonlinear and non-autonomous model of consensus formation. In: Elaydi S, Ladas G, Popenda J, Rakowski J (eds) Communications in difference equations. Gordon and Breach, Amsterdam, pp 227–236
86. Hegselmann R, Krause U (2002) Opinion dynamics and bounded confidence models, analysis and simulation. J Artif Soc Soc Simul 5: http://jasss.soc.surrey.ac.uk/5/3/2.html
87. Fortunato S (2004) Damage spreading and opinion dynamics on scale free networks. cond-mat/0405083
88. Fortunato S (2004) The Krause–Hegselmann consensus model with discrete opinions. Int J Mod Phys C 15:1021–1029
89. Fortunato S (2005) On the consensus threshold for the opinion dynamics of Krause–Hegselmann. Int J Mod Phys C 16:259–270
90. Pluchino A, Latora V, Rapisarda A (2005) Compromise and synchronization in opinion dynamics. physics/0510141
91. Ben-Naim E, Krapivsky PL, Redner S (2003) Bifurcations and patterns in compromise processes. Physica D 183:190
92. Weisbuch G, Deffuant G, Amblard F, Nadal JP (2001) Interacting agents and continuous opinions dynamics. cond-mat/0111494
93. Bonabeau E, Theraulaz G, Deneubourg JL (1995) Phase diagram of a model of self-organizing hierarchies. Physica A 217:373–392
94. Sousa AO, Stauffer D (2000) Reivestigation of self-organizing social hierarchies. Int J Mod Phys C 11:1063–1066
95. Stauffer D, Sá Martins JS (2003) Asymmetry in hierarchy model of Bonabeau et al. cond-mat/0308437
96. Schulze C, Stauffer D (2004) Phase diagram in Bonabeau social hierarchy model with individually different abilities. cond-mat/0405697
97. Malarz K, Stauffer D, Kułakowski K (2005) Bonabeau model on a fully connected graph. physics/0502118
98. Lacasa L, Luque B (2005) Bonabeau hierarchy models revisited. physics/0511105
99. Schulze C, Stauffer D, Wichmann S (2008) Birth, survival and death of languages by Monte Carlo simulation. Commun Comput Phys 3:271–294

Books and Reviews

Castellano C, Fortunato S, Loreto V (2007) Statistical physics of social dynamics. arXiv:0710 3256

Schweitzer F (ed) (2002) Modeling complexity in economic and social systems. World Scientific, Singapore

Weidlich W (1991) Physics and social science – The approach of synergetics. Phys Rep 204:1–163

Social Processes, Simulation Models of

KLAUS G. TROITZSCH
Universität Koblenz-Landau, Koblenz, Germany

Article Outline

Glossary

Social process A social process in the current context can be defined as any series of events occurring where several human beings enter into interaction. The definition may be extended to processes in which social animals are involved.

Model A model in the current context is defined as the representation of some part of the real world, which is called a target system, where this representation is done in terms of some language, be it natural language or some formal graphic and/or computer-executable language. It represents real-world entities with nouns in natural language or objects in typical programming languages, properties of these entities with adjectives in natural language or instance variables in typical programming language, actions of these entities with verbs in natural language or methods and procedures in programming languages, and relations between such entities with relations between nouns or objects in the respective languages.

Simulation Simulation is the execution of a formal model of some process described in a programming language on a computer.

Agent An agent in the current context is usually defined as a piece of software that has some autonomy, i. e. operates without other parts of a larger computer program having direct control of its internal state and actions, that can communicate with other agents in some kind of language, i. e. without direct access to other agents' inner state, that can perceive its environment and can react on it, and that can also take the initiative, i. e. can act without a stimulus from outside and display goal-directed behavior.

System A system in the current context is a mathematical structure consisting of a set of components (composition) belonging to the system, another set of things of the same kind not belonging to the system (forming the environment of the system) and a set of relations defined on the set of things forming both composition and environment and containing at least one relation that changes the history of the involved things [6]. In other contexts, "system" may have other connotations.

Definition of the Subject

The simulation of social processes has been a topic of both the social sciences and computer science for more than five decades and goes back to the first applications of computers. This is, among others, due to the fact that the first developers of computers were not only mathematicians and people from electric engineering, but also economists and even political scientists. As one can read in biographies of Herbert A. Simon, a political scientist, [61] and John von Neumann, a mathematician and game theorist, [49] persons like these were fascinated by the idea that beyond mathematical analysis computer applications could be made fruitful for the understanding of social processes.

Whereas cross-sectional analysis in the social sciences had been supported by statistical methods even earlier, any longitudinal analysis of societies needed more advanced methods of analysis than statistics can provide.

Simulation of social processes comes in different forms, as this entry will deal with in much more detail. One usually distinguishes between pure macro-level simulations where one is only interested in what happens on one level, usually a nation, a society, or an organization and where only one entity – the whole system – is represented, from different kinds of multi-level simulations where at least a system and its components or elements are represented together with the interactions between levels and interactions among the components.

Simulation of social processes has several distinct aims. On one hand, and this is usually the primary, but most primitive aim, one is interested in the prediction of processes (this is often done before the process is properly understood and often leads to predictions that can only be invalidated by empirical research performed after the prediction). On the other hand, one is first interested to understand what is going on in a social process (i. e. to communicate to oneself what one believes are the ingredients of a social process), to make this communicable to others in order to be able to discuss one's understanding with colleagues, and to experiment with the model in order to find out whether the executable simulation model behaves like the target system it represents. Simulation thus plays a similar role as earlier kinds of models, e. g. a mathematical model in physics, which were qualified by Max Planck [48] as "a system of terms and sentences, the so-called physicist's world view, which he endows to the best of his knowledge in order that it – put in place of the real world – should send him – if ever possible – the same message as the real world would send". And a computer program written for simulation is just "a system of terms and sentences" that – even literally – is able to send messages to its user.

Introduction

This entry is organized as follows. The next section tries to give an overview of social processes that have so far been modelled for the purpose of simulation. As one will see, there is a very wide range of social processes that have been simulated with quite different simulation tools and approaches. And this section will, of course, give a more detailed definition what a social process is, and it will shortly deal with some system theories (unfortunately there are very different ones under the same name) and analyze how their adherents have used simulation to analyze what they described as systems.

Section "Microanalytical Simulation" and Sect. "Multilevel Simulation" will deal with two classical approaches to social simulation. The two classical approaches are those of system dynamics and of microanalytical simulation. Both of them have a history of more than 50 years (their learned societies celebrated their 50th anniversaries in August 2007) and have found a wide audience, with system dynamics still being an approach used from ecology to management sciences, modelling ecosystems as well as the planet as a whole, but also firms and markets, and with microanalytical simulation still being an approach used to predict consequences of tax and transfer reforms and the age structure of future generations. While system dynamics is a purely one-level, one-of-a-kind approach as it always represents exactly one undivided real-world system in terms of its properties, thus hiding its components, microanalytical simulation represents a large number of non-interacting men and women and/or households that react on the stimuli given by the experimenter who is only interested in the aggregate statistics which in turn never respond to the behavior of the micro-entities. Thus, in both of these classical approaches an important feature of social processes is in a way missing, as both of them do not actually represent interactions between human beings and, moreover, do not represent interactions between individual and group or individual and society.

Section "Multi-Agent Simulation" will deal with several contemporary approaches that have not yet made their way into the ministries or the boards of enterprises (with a few important exceptions that will be discussed). Common to all of these disparate approaches is the attempt at representing processes in which interactions among individuals and between levels play a major role. Interactions can occur in topological neighborhoods (as for instance in cellular automata) or in networks (as in many agent-based or multi-agent simulations), sometimes neither neighborhood nor network topology is represented, and models restrict themselves to the micro-macro interaction. Approaches of this type most often try to represent cognitive processes beside social processes, as at least in social systems of human beings much of the interaction is controlled by cognitive processes in the minds of participants. And it is still an open question whether a valid representation of human social system can do without taking account of cognitive processes, as it is an open question whether taking only cognitive processes into account leads to an explanation of social processes ("mind is not enough" [8], but "society is not enough" seems also to be true).

Social Processes and Social Systems

Talking about social processes makes it necessary first to define what a process is. Following Bunge [5] a process is a sequence of states a thing had or has or will have over time. It can be written as a time series whose elements denote the state of a thing at a certain point of time, the sequence of states being ordered with respect to time. Thus a process can be represented with a sequence of vectors, each vector representing the state of the thing at a certain time, and each element of each vector representing one of the properties of the thing that describe its state at this time.

When talking about social processes, what was called "thing" in the preceding paragraph (and thus following Bunge [5]) will always be a system composed of human beings (and perhaps containing several levels of subsystems) where the state of the system can and should be described both on the system level and on the component level (and perhaps on the levels defined by different subsystems). Although system dynamics describes the state of the system only on this upper level and neglects any subsystems and components, and although microanalytic simulation describes the state of the system only on the level of either households or individual (but usually not on both) and not on the aggregate level (this is done only for calculating results), both of these approaches describe social processes but in an incomplete manner.

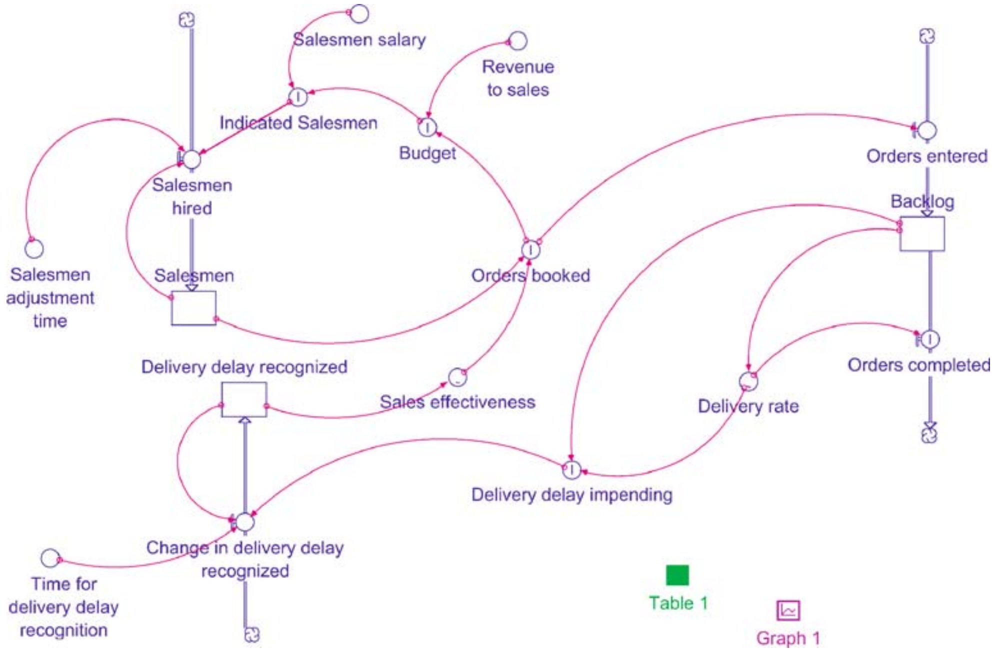
But it will not be sufficient just to describe a social process (as if videotaping it and looking at the video frame after frame), instead it is necessary for understanding and explaining why the process developed the way it developed to construct a model as a means that can create the next frame from the former frames, thus (re-) constructing the laws the process obeys. For simple things such as those elementary physics deals with this law was often represented with a differential equation, one of the classical means of modelling processes. For social systems, this representation is only rarely if ever appropriate, as the calculus does not easily lend itself to describe a population and its members at the same time. This is why system dynamics sets the members of a population (members of an organization, citizens of a country) aside and concentrates on the relatively few properties of the system as a whole, which is also done in some of the classic system theories – for Rapoport (pp. 8–9 in [52]), for instance, "the simplest mathematical description of a system – a model of a spring balance" is the "equation $L = L_0 + kW$ where L is the length of the stretched spring, W the suspended weight, L_0 the 'natural' length of the spring (when $W = 0$), and k a constant that represents the 'stretchability' of the spring." This is exactly the same view on a system as the one taken by system dynamics, where only the variable and constants attributes of the system seem to be interesting, but not the interactions between the parts of a system.

At least from the time of Coleman, but certainly even earlier, e. g. in Durckheim [14], social processes have been superficially represented with the so-called "Coleman boat" or "Coleman bath-tub", a figure [9] that mirrors the notion that processes on the level of, say, society, can be observed on this level, but it is clear that it is not society that "acts", but whatever happens on the level of society is performed by the actions of individual human beings that are in a way under the control of the society and at the same time making up society. In Coleman's example, there is an observable process in which protestant religious doctrine seems to have led to capitalism. This "single proposition breaks into three: one with an independent variable characterizing the society and a dependent variable characterizing the individual; a second with both independent and dependent variable characterizing the individual; and a third with the independent variable characterizing the individual and the dependent variable characterizing the society. Thus the proposition begins and ends in the macro level, but in between it dips to the level of the individual" [8]. The original macro proposition is usually represented with the deck of a boat, while the first of the three new proposition forms its stern, the second is the keel and the third is the bow of the boat. Thus, protestant religious doctrine introduced in a population (wherever this doctrine might originate) influences the value system of individuals, and the changed value system modifies their economic behavior; finally this economic behavior of the individuals modifies the characteristics of the population as a whole (and this latter process is often called "emergence").

The first of these inter-level processes is meanwhile known as "immergence", a term coined by Castelfranchi [7] for the process by which – in Colemans's example – an individual becomes aware of a doctrine ("cognitive emergence") and decides to act according to this doctrine; this process was already described by Durckheim [14] when he talks about "sociological phenomena [that] penetrate into us by force or at the very least by bearing down more or less heavily upon us". The other process, by which individuals change the state of the system (population, organization, ...) qualitatively is often called "emergence".

System Dynamics

As superficially mentioned in earlier sections, system dynamics is a tool for modelling processes in social systems on a macro level. The approach originated in the 1950s, when Jay W. Forrester made his first steps in representing social processes in terms of differential equations (as others had done before him, e. g. Herbert A. Simon [59,60]) and found it necessary to include non-differentiable functions into his models. This was how the programming language DYNAMO was first invented, meanwhile superseded by more modern attempts at designing modelling languages with graphical user interfaces, a functional, i. e. declarative programming language, in which the programmer does not (have to) describe how the computer should perform its calculations, but only



Social Processes, Simulation Models of, Figure 1
Graphical representation of Forrester's sales growth model, using the STELLA simulation tool

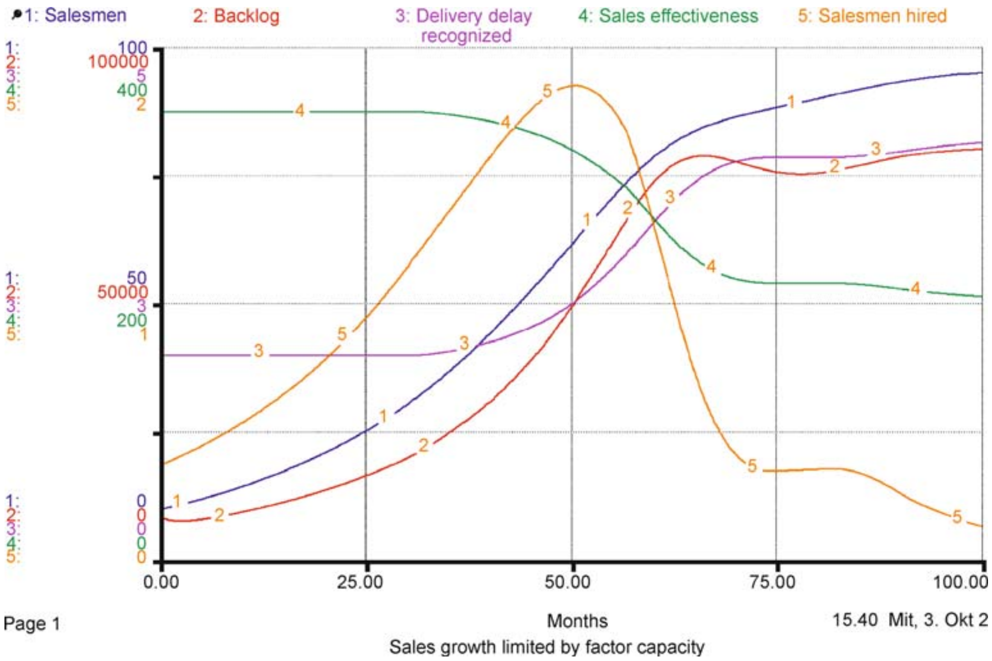
which invariants should hold over time. An invariant in this sense is an equation that holds true even when the states of a thing or system change (for instance, even if the size of population is changing over time, its growth rate could be constant for a long period, such that for all times the equation $\text{size}(t+1) = \text{size}(t) + \text{growthrate} * \text{size}(t)$ holds). System dynamics is of course much more sophisticated, but the equation in the preceding sentence is a typical representative of equations in system dynamics models. The first step in modelling is usually a graphical model showing the dependencies and feedbacks between the attributes of the modelled macro system. These attributes come in several different kinds, the two main kinds of variables being levels and rates. Levels are state variables, usually describing something like material stocks (inventory of a warehouse, size of a population etc.), whereas rates are state change variables, usually describing the material flow into a stock and from a stock per time unit. Beside these, auxiliary variables serve as abbreviations, and constants can be defined.

Among the target systems modelled with the help of system dynamics and the respective tools, there are agglomerations [18], industrial enterprises [16], ecosystems [2] as well as the world as a whole [19,42,43].

A relatively simple example of system dynamics modelling can be found in [17] where a firm with its "sales growth and saturation" is modelled and simulated. The only properties of this firm considered are the current salesperson staff, the backlog of unprocessed orders and the delivery delay recognized. These "level variables" are connected to each other via several "rate variables" – salespersons hired, change in delivery delay recognized, orders entered and orders completed – and several "auxiliary variables" – sales effectiveness, orders booked, budget, indicated salespersons, delivery rate, delivery delay impending – and a few constants – salespersons adjustment time, revenue to sales and time for delivery delay recognition. A STELLA representation of the model makes the idea behind the model sufficiently clear (see Fig. 1).

The simulation of this model shows the time series in several variables of this fictitious firm: The model starts with 10 salespersons and a high budget (stemming from a large number of orders), thus more salespersons are hired who bring more and more orders, more than can be processed such that the backlog quickly increases and the sales effectiveness decreases after some time. This is the signal for hiring less new salespersons, but the delivery delay cannot be decreased in the long run, as a still rising number of salespersons continue to bring more orders.

The graph in Fig. 2 shows some of the drawbacks of the approach: at any time (except at the start) the number of salespersons is a non-integer number, and at any time the number of newly hired persons is also a non-integer



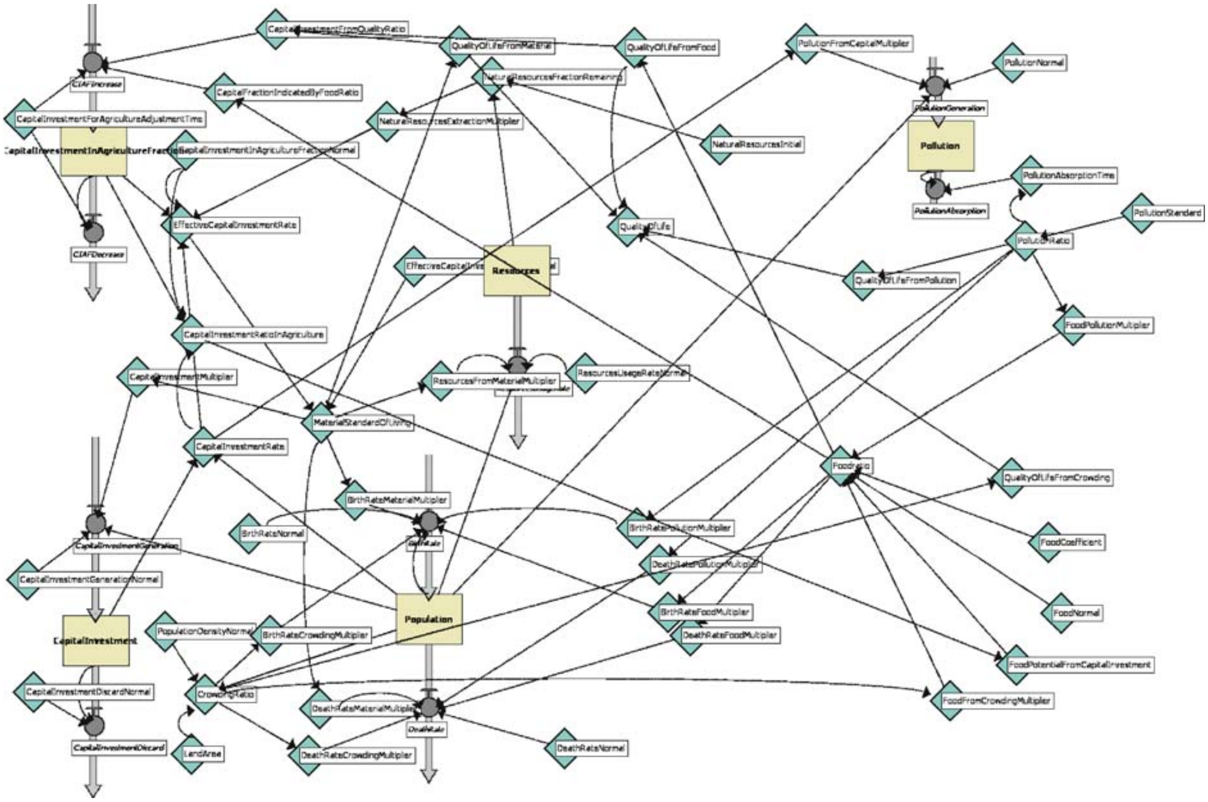
Social Processes, Simulation Models of, Figure 2
Graphical output of the sales growth STELLA model

number ranging from 0.13 to 1.85 whereas, of course, in a real firm, persons would be hired on a full-time or half-time base, but numbers such as 0.13 or 1.85 would be quite unlikely.

At the other end of the range of applications, the world models of the early 1970 can be placed, which became very famous as the Club of Rome funded and published them. Their objective was to predict the consequences of the world population growth for the use of resources and the pollution. Forrester's version [19] is replicated in a NetLogo 4.0 reimplementation shown in Fig. 3. This model has only five level variables – population, resources, pollution, capital investment (in industry) and capital investment in agriculture – which are linked together by nine rate variables and a large number of auxiliary variables and constants. The growth of the world population, for instance, is determined by two rate variables that contain the numbers of births and deaths, respectively, per time unit, and these in turn are determined by natural birth and death rates (biological constants, as it were) and several multipliers, auxiliary variables that transport the influences of causes such as pollution, food availability etc.; resources is another level variable, and this one is only determined by one rate variable, namely the decrease in resources per time unit, as Forrester (and also his successors in Meadows's research group) did not foresee that new resources might be discovered that were not known at the beginning of the simulation run (1900) or at the time when the simulation model was designed (1970).

The output of the model (see Fig. 4) shows what was predicted in the 1970s: a continued increase in the world population up to 8.6 billion in 2040, a decrease in available resources from 900 billion units (whatever these mean) in 1900 to 445 billion in 2040, and an increase in pollution from 200 million to 20 billion units (of another kind) in about 2050 – a result that was discussed worldwide in the 1970 (and perhaps this discussion was the reason why the predictions did not come out true 40 years later – although Meadows in [43] seems to have believed that his predictions were validated by the two decades after [42] appeared.

It is one of the big advantages of system dynamics models that they are easy to understand, easy to develop and to maintain and that they deliver plausible predictions that can easily be validated (or invalidated) against empirical data, though only after several decades. One of their important drawbacks is that the estimation of their input parameters is difficult if not impossible. Take as an example the estimation of a function between the pollution ratio and the pollution dependent death-rate multiplier (the pollution ratio is the actual pollution divided by the "pollution standard", which is defined as 3.6 billion pollution units – approximately the pollution in 1975). In Forrester's world model this function is programmed as


CapitalInvestmentFromQualityRatio
QualityOfLifeFromMaterial
QualityOfLifeFromFood
PollutionFromCapitalMultiplier
PollutionNormal
CIAFIncrease
CapitalFractionIndicatedByFoodRatio
NaturalResourcesFractionRemaining
CapitalInvestmentForAgricultureAdjustmentTime
NaturalResourcesExtractionMultiplier
NaturalResourcesInitial
PollutionGeneration
Pollution
CapitalInvestmentInAgricultureFraction
CapitalInvestmentInAgricultureFractionNormal
PollutionAbsorptionTime
PollutionAbsorption
PollutionStandard
EffectiveCapitalInvestmentRate
QualityOfLife
CIAFDecrease
PollutionRatio
EffectiveCapitalInvestmentRatioInAgricul
Resources
QualityOfLifeFromPollution
FoodPollutionMultiplier
CapitalInvestmentRatioInAgriculture
CapitalInvestmentMultiplier
ResourcesFromMaterialMultiplier
ResourcesUsageRateNormal
MaterialStandardOfLiving
CapitalInvestmentRate
Foodratio
QualityOfLifeFromCrowding
BirthRateMaterialMultiplier
BirthRateNormal
BirthRatePollutionMultiplier
CapitalInvestmentGeneration
BirthRate
FoodCoefficient
CapitalInvestmentGenerationNormal
DeathRatePollutionMultiplier
Population
BirthRateFoodMultiplier
FoodNormal
CapitalInvestment
PopulationDensityNormal
BirthRateCrowdingMultiplier
DeathRateFoodMultiplier
FoodPotentialFromCapitalInvestment
CapitalInvestmentDiscardNormal
CrowdingRatio
FoodFromCrowdingMultiplier
DeathRateMaterialMultiplier
DeathRate
CapitalInvestmentDiscard
LandArea
DeathRateCrowdingMultiplier
DeathRateNormal

Social Processes, Simulation Models of, Figure 3
Forrester's WORLD2 model in a NetLogo 4.0 implementation

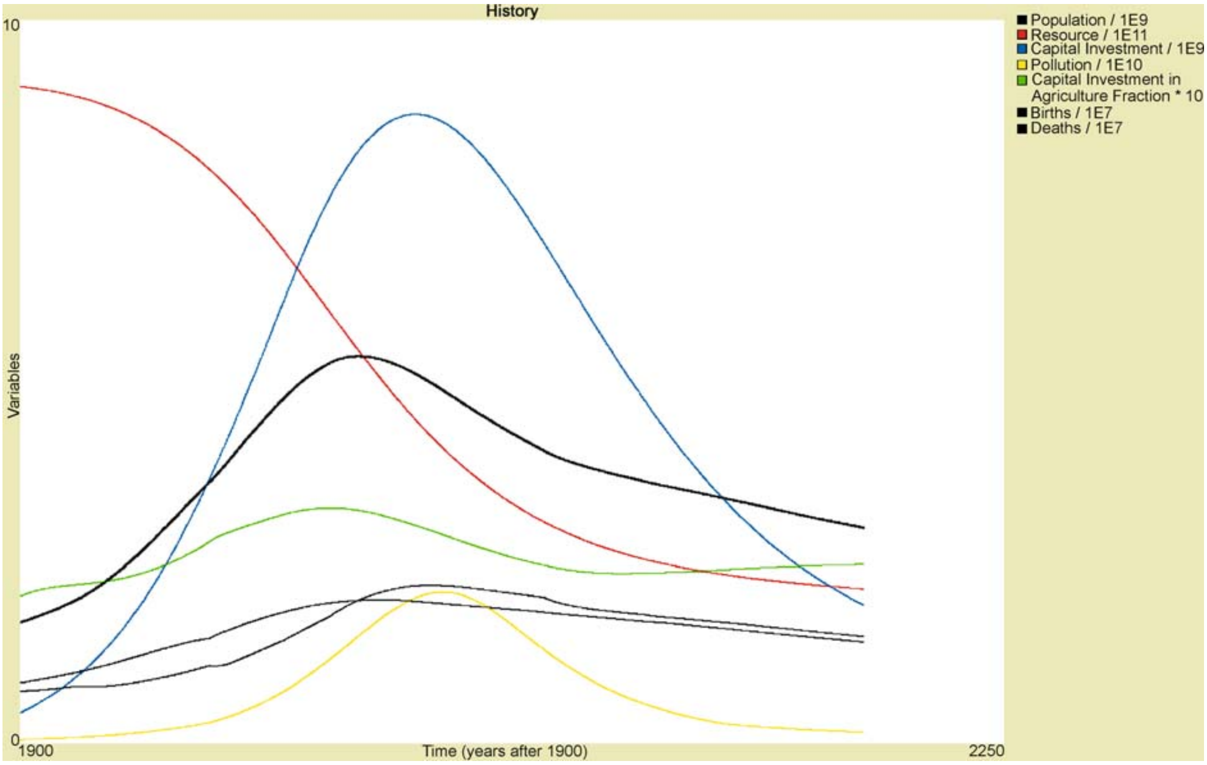
a table function (see Fig. 5), but in many cases it is quite questionable how the form of such a function was estimated.

In [42], the estimation of parameters for many functions between auxiliary variables is explained. The strategy of estimating the parameters of these functions is in most cases the same: it is usually the result of a nonlinear regression between – to take the preceding example – the pollution rate and the death rate pollution multiplier estimated from as many countries of the world as data could be found from, thus believing that synchronously found data from various places in the world were representative of data for the world as a whole collected at different points of time.

Microanalytical Simulation

Microanalytical simulation [22,47] has its origin in the necessity of predicting tax income and tax loads as well as the cost of social transfer. Every modern state needs information on how much taxes must be levied on its citizens in order to perform all the tasks of the state and how much social security contributions must be levied in order to pay retirement income, and it needs to know how many children will enter and leave the school system, how many students will enter and leave universities, how many people will enter the labor force at what time of their life and leave it at the time of their retirement, and, lastly, how many people will have near relatives who might be ready and willing to nurse them when they are in need. Whereas the age structure can be estimated even with difference equations, thus with system dynamics techniques, at least the last of the mentioned questions (which takes kinship networks into account) can only be answered on the micro level.

Microanalytical simulation always starts with a large representative sample of the target population. The sampling units are usually households (of which all members are interviewed), and the attributes sampled are at least age, gender, marital, educational and socio-economic status as well as the links to other members of the same household. This sample is then updated for every time step (usually one year) to answer the questions mentioned above for some time in the future.



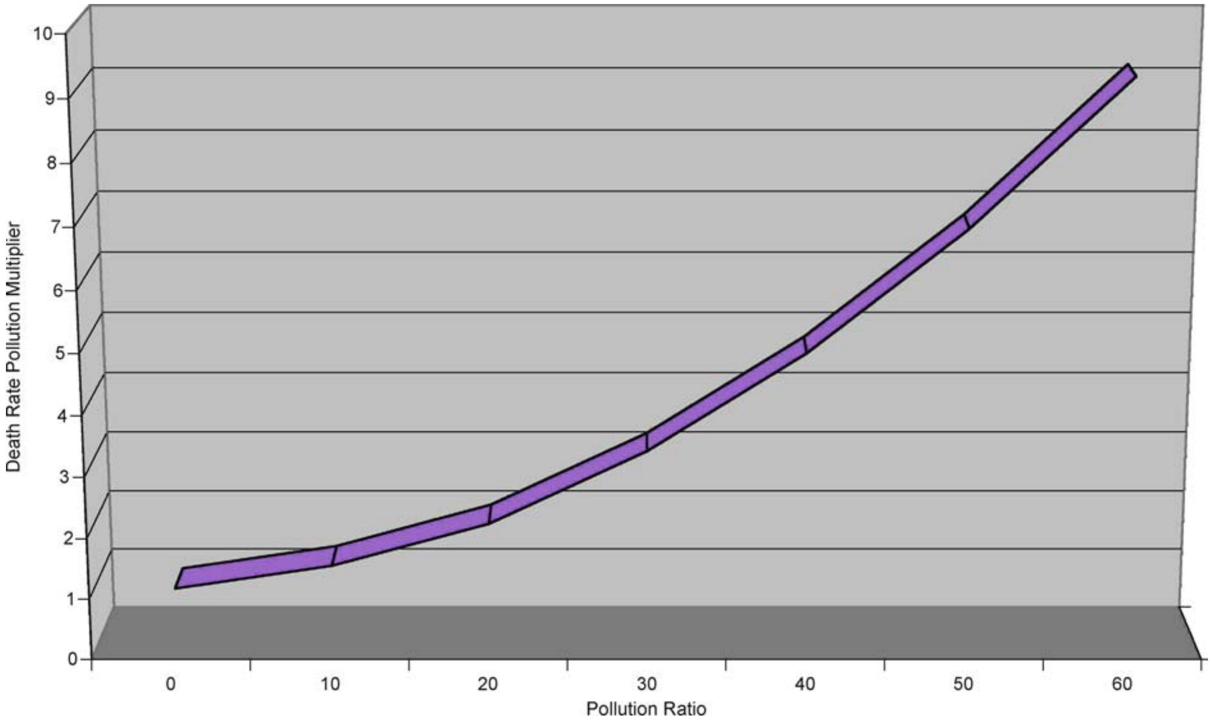
Social Processes, Simulation Models of, Figure 4
Some time series of Forrester's WORLD2 model, from a NetLogo 4.0 implementation

Microanalytical simulation has two easily distinguishable variants. They are usually labelled "static microsimulation" and "dynamic microsimulation". The former is static in so far as the attributes of the sample individuals are not changed over time, only the weights of the individual are updated in order to be representative for some macro variables where the information about the composition are taken from other (demographic or econometric) models. This static simulation is used only for short-time predictions, for instance for estimating tax loads due to revised tax regulation. Dynamic microsimulation changes all individual attributes according to known or estimated probabilities of state changes (for instance age-dependent birth, death, marriage and divorce rates or frequencies of transitions within the educational system and from the educational system into the labor force). For every time step, every micro entity has to be updated, and its link to other micro entities has to be revised, too. As compared to static microsimulation, this variant has the advantage that all state changes are endogenous and, more important, that state transition probabilities can be disaggregated to small sectors of the target and sample population (for instance, assuming different probabilities of giving birth to a first, second, third etc. child for women of different age groups, of different educational and socio-economic state), given that sufficient evidence for estimating such probabilities is available.

Early microanalytic simulation models were always programmed in general purpose programming languages from scratch, without using any particular tools. Their results were only partially published as research was mostly done for state agencies who were not always interested in having all the details of models published. Only recently, toolboxes such as UMDBS [56] and CoMicS [31] became available which lend themselves to be used in academic research and teaching, the more so as in many countries data sets of large representative samples and even panels have become available to the academic public for more than a decade now.

Multilevel Simulation

Multilevel and multi-agent simulations, different as they may be, always had their focus on interactions either among agents of the same kind or between agents of different levels or both. With levels, these approaches under-



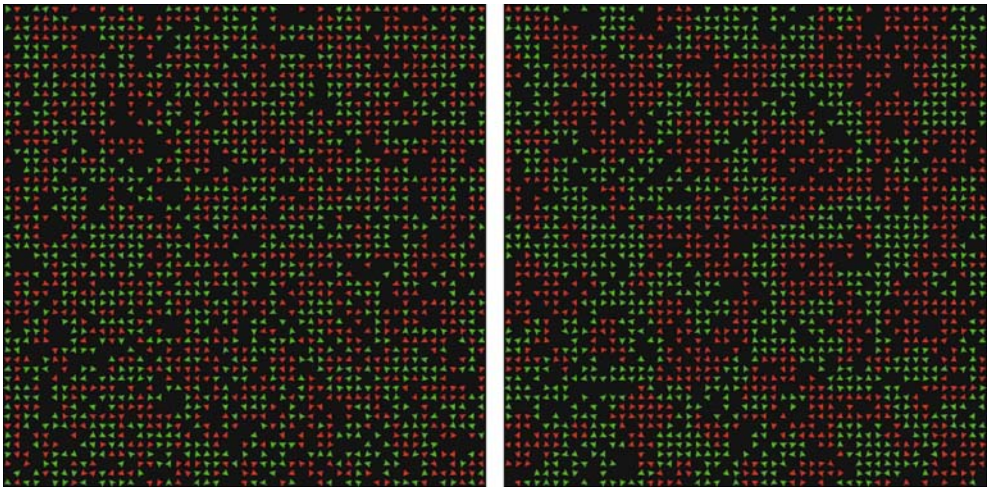
Social Processes, Simulation Models of, Figure 5
How system dynamics defines functions between its variables: the dependence of the death rate multiplier on the pollution ratio

stand different kinds of systems, where the elements of the systems of one kind are systems of the other kind [6]. To give an example: From the biologist's point of view, human beings are systems of one kind (composed of different kinds of tissues, or of cells, etc.), and on the other hand they can be seen as elements of systems called human groups, or populations, or societies etc. And what multilevel simulations are interested in is the interaction between, for instance, the group and an individual, between the macro and the micro level, thus focusing on the stern and bow of the "Coleman boat" [9]. Multi-agent simulation models are usually less focused on the inter-level interaction and more on the inter-agent interaction, but they, too, are usually interested in detecting properties emerging on the macro level.

More often than not, the label "multi1-agent simulation" is used for models in which the micro entities display only very simple behavior, have only very few states and do not communicate with other micro entities (in any reasonable sense of the word "communicate"), have no memory and make no decision. To distinguish these from multi-agent models proper (see the next section), they will here be labelled as "multilevel" models.

A first, quite well-known example may illustrate this. Thomas Schelling [57,58] designed a model in which (rather primitive) agents have to decide whether they move from one neighborhood into another, in which they can expect to be happier. Their happiness depends on the composition of their neighborhood, given that agents differ only with respect to one dichotomous property (say, their native language) – they are the happier, the more people in their neighborhood speak the same language, and there is a threshold where happiness turns into unhappiness. The world in Schelling's model is represented with a cellular automaton (see the chapter on cellular automata in this Encyclopaedia), and the neighborhood consists of the eight cells around an agent, which may be empty or inhabited by another agent. The central agent is happy if the language distribution in its neighborhood is such that a certain percentage of its neighbors speak the same language. In each time step, every agent scans its neighborhood and decides whether to move or stay, if it decides to move it tries to find another cell where the neighborhood is more attractive than in its original cell. One realization [67] of this model is shown in the two parts of Fig. 6.

It shows that the random distribution of red and green agents soon changes into a distribution with clearly distinguishable red and green clusters, part of which are even separated by uninhabited zones. The threshold between



Social Processes, Simulation Models of, Figure 6

Schelling's segregation model (*left:* initialization, *right:* state of the macro level when all individuals are happy with their neighborhoods)

happiness and unhappiness in this example is 30 per cent, i. e. an agent moves only if more than 70 per cent in its neighborhood are different from itself. Although these agents are quite tolerant with respect to strangers, clusters occur; and these clusters are larger and better separated when the tolerance threshold is higher. One should note here that the Schelling model takes only local interactions between agents into account; there is no interaction between levels if only in so far as the structure of the macro level changes due to the individual interactions, but it does not feedback into the micro level as the agents are not capable of recognizing what happens on the macro level.

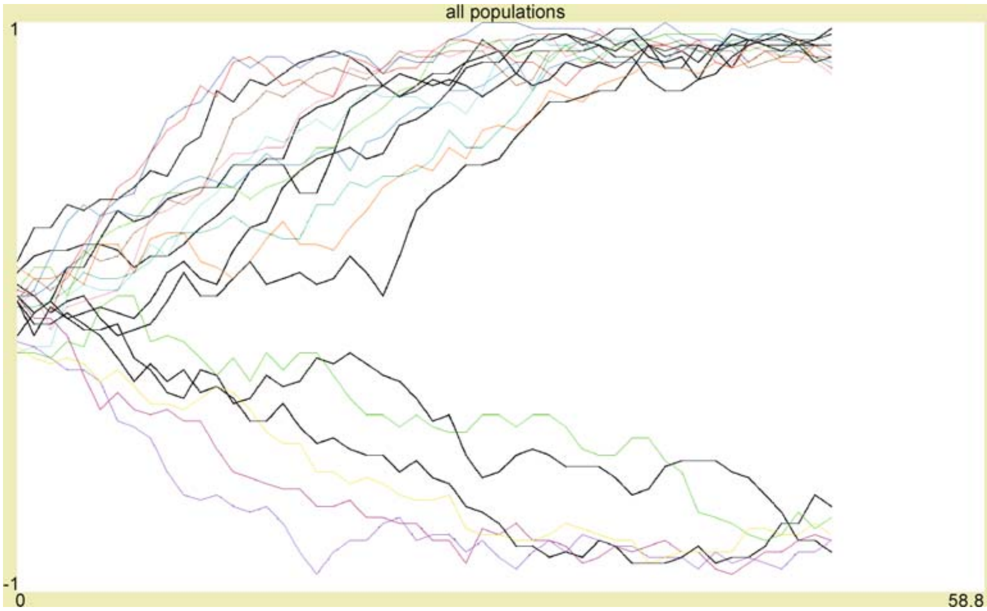
Another early approach to multilevel modelling of social processes is mainly interested in the interactions between the levels. It was originally developed by Hermann Haken [21], a German physicist who tried to apply the methods with which he had done research in laser physics to other disciplines, including sociology, and extended by his colleagues Wolfgang Weidlich and Günter Haag [66]. This approach, called synergetics by Haken, is currently better known as sociophysics or econophysics. Although the micro level entities are often also called agents by several authors, they are only particles as they only react on something like social forces or fields that are modelled in a manner very similar to the description of gravitational, electrostatic and electromagnetic forces and fields.

The original work by Weidlich and Haag [66] describes populations consisting of a large number of individuals where the individuals usually have only one discrete state variable which is either binary or can take only very few values, and the state change of the individuals depends on the macro state which is in turn determined by the individual state changes. The simplest example may make this clear: Imagine a population whose members have to decide between two options (yes and no) and whose decision depends on the perceived current majority in such a way that the probability of switching from no to yes is the higher, the larger the current yes majority is and the other way round. Then under certain assumptions of modelling the exact function between the current yes majority and the individual transition probabilities, the population as a whole shows some interesting behavior. In the example the two individual transition probabilities have the following form:

$$\mu_{\text{yes}\leftarrow\text{no}} = \nu \exp(\pi + \kappa x)$$

$$\mu_{\text{no}\leftarrow\text{yes}} = \nu \exp[-(\pi + \kappa x)] \,,$$

where ν is a flexibility parameter, π is a preference parameter, and κ is a coupling parameter, whereas x describes the current state of the population ($x = -1$ means "all no", $x = +1$ means "all yes"). The flexibility parameter only determines how likely any change between options is; the preference parameter determines whether changes from yes to no are generally more likely than those from no to yes (if it is negative) or the other way round (if it is positive); and the coupling parameter determines the strength of the influence the population exerts on the individual. It goes without saying that any individual change changes the state of the population. Figure 7 shows a typical outcome of a simulation of this very abstract social process. Each of the curves represents the history of the state of one out of twenty populations (all of them with a vanishing preference parameter and a coupling parameter of 1.4);



Social Processes, Simulation Models of, Figure 7
History of 20 populations whose members change their state with a probability that depends on the state of the population

some of the curves show that their populations soon develop into populations with a very high yes majority while others become populations with a high no probability, but none of them is ever entirely homogeneous. A more concrete social process that is likely to behave similarly is the distribution of two alternative goods that are not easily compatible, e. g. computer operating systems or videotape machines; here one could expect a high coupling as being with the majority makes the exchange of programs or videos easier. A low coupling constant results in an outcome where all populations are more or less evenly split in yes and no members – which in reality could be the case when the two alternative goods are easily compatible or substitutable, e. g. cigarette brands.

Extensions of this model were proposed by Lumsden and Wilson [41] who replaced the exponential function in the formula for the individual transition probabilities with other types of functions, to explain the co-evolution between "genes, mind, and culture". Other extensions, already given in [66], deal with several populations interacting with each other, the individual transition probabilities not only depending on the state of the members' own population, but also on the state of other populations, for instance modelling migration behavior of two populations between two regions.

Helbing [26] extended the models of his teachers, Wolfgang Weidlich and Günter Haag, in a way that he described as "quantitative sociodynamics". Much like his teachers he used closed solutions wherever possible to analyze the abstract social processes that he formalized, at the same time extending the focus not only to interactions between levels but also to interactions between individuals [24,25] which he also used to model the behavior of traffic participants, mainly pedestrians [24,27]. Here again, individual human beings are modelled as particles in a field and not endowed with the familiar human capabilities of decision making. In Helbing's pedestrian models (e. g. [27]) the pedestrians are just particles moving in a "social field", a term originally coined by Lewin [38], according to whose theory "behavioral changes are guided by so-called social fields or social forces, which has later on been put into mathematical terms" (p. 625 in [27]). These particle-pedestrians are accelerated at any time by a force which is determined by their desired directions and speeds and a repulsive force which keeps them apart from each other and from obstacles. Although the comparison between simulation results and video recordings of real pedestrian behavior shows nearly no qualitative difference, the particle-pedestrian are far from the definition of agents that is used by other authors who describe their simulations as multi-agent models.

Another kind of "social field" models was used to analyze attitude formation processes in an artificial society where the attitudes were not binary or categorical, but continuous. Models like these were inspired by Anthony Downs's early theory on party affiliation [13]. Downs assumed that voters moved in a continuous attitude space (one-dimensional, from left to right) and were attracted or repelled by parties, and he believed that under certain circumstances a polarization of the electorate would emerge with high frequencies on the left and right wings and low frequencies in the middle, but he never formalized this theory. Several different formalizations were published in the sequel which tried to explain polarization without the interference of parties. One of these formalizations was inspired by empirical data which – for some German elections in the past – showed such a polarization in terms of a bi- or multimodal frequency distribution of voters over their attitude space which emerged during election campaigns and vanished shortly after the election [63]. The simulation runs started with a sample of a large numbers of model entities which obeyed an approximate bivariate normal distribution at the beginning (samples are never perfectly normal) entities then tried to move according to a function which was composed of the gradient of the frequency density distribution and some random effect. These movements changed the frequency density distribution, which in turn changed the gradient. In most simulation runs the frequency density functions became bimodal or multimodal after some time. Another representative of models inspired by Downs was presented by Guillaume Deffuant and is colleagues [12]. Here the simulated entities have a one-dimensional continuous attitude (much like in [63]), but additionally something like persuasive power, and as in [63], typically multimodal distributions emerge with modes both in the center but also at the ends of the attitude range, representing extremists in real electorates.

The approach in [12] is in some way similar to social impact theory developed by Bibb Latané [34] who also combined attitude and persuasive power, but let simulated entities interact on a grid, thus introducing topography and restricting interaction to near-neighbor interaction. The simulation models following his theory [35,36] show, similarly to the synergetic models, a survival of minorities, but also the clusters observed in Schelling's migration models (although the simulated entities in the social impact theory do not move in topological space).

Multi-Agent Simulation

An example that can easily be compared to the pedestrian models is the riot behavior model of Wander Jager and colleagues [29]. Here, agents belong to two different subpopulations (party 1, party 2) which can be imagined as the fans of two competing football clubs. Besides moving in a two-dimensional space, they can decide to fight or withdraw. In a version with police as a third subpopulation, the agents of this kind have additional behavioral options. Which action an agent takes does not depend on anything like a social field (at least not in the sense of Helbing, but perhaps in the sense of Lewin), but on their perceptions of the behavior of other agents in their neighborhood. These agents follow more or less the definition of agents that was given by Wooldridge and Jenkins [69] and which is currently the most often cited definition of agents:

An agent is a piece of software that

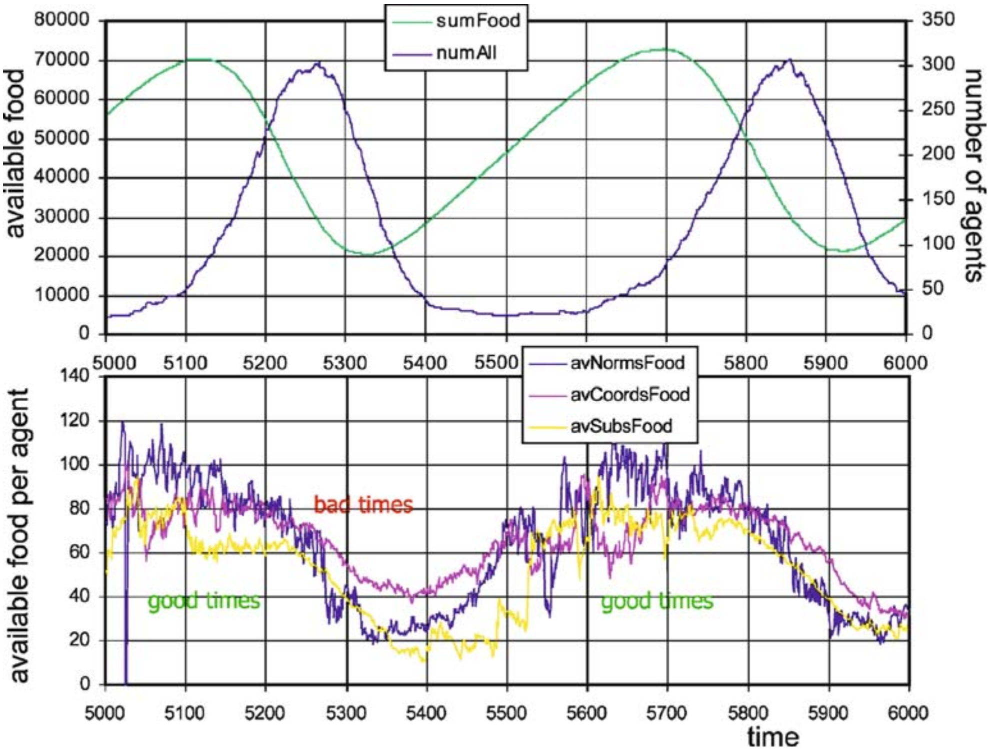
- has some autonomy, i. e. operates without other parts of a larger computer program having direct control of its internal state and actions,
- can communicate with other agents in some kind of language, i. e. without direct access to other agents' inner state,
- can perceive its environment and can react on it, and
- can also take the initiative, i. e. can act without a stimulus from outside and display goal-directed behavior.

This results in several requirements: In order to be able to take the initiative (to be pro-active), agents must have goals (descriptions of desired and possible states of their neighborhood or surroundings, including themselves) and must be able to compare their perceived state of their environment to their goals. According to the outcome of such a comparison, they must decide between several possible actions which they have to evaluate, and moreover they must be able to design a plan (i. e. a course of possible actions) if there is no action that leads them directly to their goal. Another requirement can be that agents communicate with other agents, for instance asking them for their help when only co-operation enables them to achieve their goals.

Among the early forerunners of multi-agent systems in the social sciences at least two can be named where processes of voter attitude changes are modelled and simulated. Although the poor computer languages of the late 1950s and early 1960s did not allow for agents in the sense of our days, any re-implementation would nowadays be a multi-agent system with several classes of agents (representing voters, candidates, media channels, as in [1] or in the Simulmatics project supporting John F. Kennedy's election campaign [11]) as they dealt with the communications among citizens, between citizens and candidates as well as between citizens and media channels and modelled their behavior and actions in a rule-based manner.

A classical example that fulfills most of the requirements raised in [69] is Epstein's and Axtell's attempt at building "Social Science from the Bottom Up" [15], an artificial world in which agents make their living on two different crops which they harvest, store, consume and barter according to their needs and the available resources. An extension of this artificial world can be found in [32]. In addition to all the features of the artificial society in [15], these agents can subordinate to others if the latter offer themselves as co-ordinators and promise their subordinates to inform them about the state of a wider environment than they can perceive individually, while the subordinates in turn provide their chieftains with their local information and pay a certain tribute from what they harvest. To be able to do all this, all agents have a long-term memory in which they store information about their world and which can guide them to places in their world that promise a rich harvest of the kind of crop they actually need. This extended model shows that a world with co-ordinators and subordinates is more sustainable than a world in which all agents try to harvest on their own; it also shows that lonely agents are best off when times are getting better, whereas subordinating pays best when times are getting worse and being a co-ordinator pays best when times are worst. Figure 8 is a part of two plots written by the simulation program: the green curve in the upper graph shows the development of food available over time, the blue one shows the number of surviving agents over time, while the three curves in the lower graph show the average "income" of the three types of agents (blue. the lonely agents, magenta: the chieftains, yellow: the subordinates) – "bad times" is when the overall available food is at its minimum, "good times" is when it is at a maximum. Among others, Fig. 8 shows that the macro behavior of the process is the same as the behavior of the well-known Lotka–Volterra model [39,65], but the classical coupled differential equations of this model or the stochastic process described in [66] would not have been able to display the additional information about the fate of the three kinds of agents.

Validating a model [64] such as the one described above is nearly impossible, as the model describes an abstract social process which is very unlikely to have ever occurred in reality. Nevertheless, this model makes a prediction that is quite similar to the prediction of another agent-based simulation [52,54]: "In a society of herdsmen and farmers in Western Africa, decisions which rest on



Social Processes, Simulation Models of, Figure 8
Development of an artificial society whose members can declare themselves co-ordinators and subordinates, respectively (graphs produced by the simulation model described in [32])

friendship networks ('friend-priority' decisions) proved to be much more effective then decisions which were made on pure cost deliberations ('cost priority' decisions)." See p. 189 in [52]. The papers quoted here are at the same time representatives of a style of modelling and simulating social processes which is known as "participative simulation" and usually used to discuss models with stakeholders who work with the simulation models and give hints at improving and amending models to adapt them better to their personal experience. This style of modelling and simulation is being used more and more often in consultancy to give stakeholders an opportunity to work with models on their own, instead having to believe what the consultant presents in the end.

Tools

Computer simulation always needs some tools to be run. In the early times of computer simulation, simulation programs were written in general purpose programming languages (like FORTRAN), but over the decades special simulation toolboxes were designed and implemented, first by researchers in fields such as organization science, production management (particularly for discrete event and queuing systems) or physics and engineering (particularly for numerical solutions of differential equations). Both of these kinds of tools were also used by social scientists, though only rarely. DYNAMO was the first simulation tool designed for and used by economists and social scientists [50], but only for system dynamics models. It was later on superseded by STELLA [53] with the same purpose, but a graphical user interface (see Figs. 1 and 2). There are several simulation tools for cellular automata (see the entry on cellular automata for more details). Multi-level simulation models are often written in general purpose programming languages (but see [44] describing a language – MIMOSE – explicitly designed for this kind of models). Multi-agent simulation proper has been developed with the help of toolboxes such as SWARM [30], RePAST [45,46], MASON [40], CORMAS [37] and NetLogo [68] (the latter was also used to produce the majority of graphics in this entry). All of these are freeware and capable of supporting most social simulation techniques (with the exception of microanalytic simulation, as this always needs a large database connected to the simulation tool). With the exception of NetLogo, which has a simulation language of its own, the actual description of agent behavior has to be done in object-oriented languages such as Smalltalk, Objective C or Java (for more detailed reviews of these toolboxes see [51,62]).

Future Directions

In a way, "agents cover all the world" [4] in that multiagent systems can be used for all simulation purposes, as agents can always be programmed in a way that they behave as continuous or discrete models, can be activated according to event scheduling or synchronously or in a round-robin manner, can use rule bases as well as stochastic state transitions, and all these kinds of agents can even be nested into each other, thus supporting a wider range of applications than any of the classical simulation approaches. This leads to a third aspect of complexity (after the complexity of domains and the complexity of time): agent-based models can encompass several different approaches, both from a technical and implementation point of view, but also from the disciplines making use of simulation (for instance, disciplines such as neurophysiology, cognitive psychology, social psychology and sociology can combine their contributions into a deeply structured simulation model).

It goes too far to say that all this would not be possible in a non-agents world (as everything is programmable in Assembler), but the examples above will have made clear that models combining aspects from even neighboring disciplines as the ones enumerated in this entry are only understandable and communicable when they come in a modular form that is typical for agent-based models – and the same applies to ecological models where disciplines from physics and biochemistry to population biology and economics would play their co-operative roles.

Although agent communication languages (see the special issue of Autonomous Agents and Multi-Agent Systems, vol. 14 no. 2) such as KQML [33] have been developed for a long time (back to 1993), it is still an open question how agents in a simulation model could develop a communication means on their own and/or extend their communication tool to be able to refer to a changing environment (see the special issue of Autonomous Agents and Multi-Agent Systems, vol. 14 no. 1). Hutchins and Hazlehurst [28] made a first step into the field of the emergence of a lexicon, but their agents were only able to agree on names of things (patterns) they saw. The NEWTIES project [20], ambitious as it is, aims at creating an artificial society that develops its own culture and will also need to define agent capabilities that allow them to develop something like a language although it is still questionable whether the experimenters will be able to understand what their artificial agents talk about. The same objective is aimed at in the current EMIL project (Emergence in the Loop [3,10]) which attempts at creating an agent society in which norms emerge as agents observe each other and draw conclusions about which behavioral feature is desirable and which is misdemeanor in the eyes of other agents – which, as in the case of language emergence, makes it necessary that agents can make abstractions and generalizations from what they observe in order that ambiguities are resolved.

This means that multi-agent models in the proper sense – where software agents model cognitive capacities of human beings – will be the aim of any future development in the simulation of social processes, although even these models will still contain features of the classical approaches, as even human beings display behavior that can be categorized as reflex acts and instinctive and which can be modelled with models borrowed from particle kinematics in which human beings are represented by stochastic automata. On the other hand it should be clear that only part of human behavior in very large groups can be explained this way, as an important part of this behavior consists in deliberate and conscious actions that are determined by cognition.

Bibliography

Primary Literature

1. Abelson RP, Bernstein A (1963) A Computer Simulation of Community Referendum Controversies. Pub Opin Q 27(1):93–122
2. Anderson JM (1973) The Eutrophication of Lakes. In: Meadows D, Meadows D (eds) Toward Global Equilibrium. Wright Allen, Cambridge, pp 171–140
3. Andrighetto G, Campenni M, Conte R, Paolucci M (2007) On the Immergence of Norms: a Normative Agent Architecture. In: Proceedings of AAAI Symposium, Social and Organizational Aspects of Intelligence, Washington
4. Brassel KH, Möhring M, Schumacher E, Troitzsch KG (1997) Can Agents Cover All the World? In: Conte R, Hegselmann R, Terna P (eds) Simulating Social Phenomena. Springer, Berlin, pp 55–72
5. Bunge M (1977) Ontology I: The Furniture of the World. Treatise of Basic Philisophy, vol 3. Reidel, Dordrecht
6. Bunge M (1979) Ontology II: A World of Systems. Treatise of Basic Philosophy, vol 4. Reidel, Dordrecht
7. Castelfranchi C (1998) Simulating with Cognitive Agents: The Importance of Cognitive Emergence. In: Sichman JS, Conte R, Gilbert N (eds) Multi-Agent Systems and Agent-Based Simulation. Springer, Berlin, pp 26–44
8. Castelfranchi C, Conte R (1994) Mind is not Enough. The Precognitive Bases of Social Interaction. In: Doran JE, Gilbert N (eds) Simulating Societies: The Computer Simulation of Social Phenomena. UCL Press, London, pp 267–286
9. Coleman JS (1990) Foundations of Social Theory. The Belknap Press of Harvard University Press, Cambridge
10. Conte R, Andrighetto G, Campenni M, Paolucci M (2007) Emergent and Immergent Effects in Complex Social Systems. In: Proceedings of AAAI Symposium, Social and Organizational Aspects of Intelligence, Washington
11. de Sola Pool I, Abelson RP (1962) The Simulmatics Project. Pub Opin Q 25:167–183
12. Deffuant G, Amblard F, Weisbuch G, Faure T (2002) How can Extremism Prevail? A study based on the relative agreement interaction model. J Artif Soc Soc Simul 5(4) http://www.soc.surrey.ac.uk/JASSS/5/4/1.html
13. Downs A (1957) An Economic Theory of Democracy. Harper, New York
14. Durckheim E (1982) The rules of the sociological method (trans: Halls WD). The Free Press, New York
15. Epstein JM, Axtell R (1996) Growing Artificial Societies – Social Science from the Bottom Up. MIT Press, Cambridge
16. Forrester JW (1961) Industrial Dynamics. MIT Press, Cambridge
17. Forrester JW (1968) Principles of Systems. MIT Press, Cambridge
18. Forrester JW (1969) Urban Dynamics. MIT Press, Cambridge
19. Forrester JW (1971) World Dynamics. MIT Press, Cambridge
20. Gilbert N, den Besten M, Botovics A, Craenen BGW, Divina F, Eiben AE, Griffioen R, Hévízi G, Lõrincz A, Paechter B, Schuster S, Schut M,C Tzolov C, Vogt P, Yang Lu (2006) Emerging Artificial Societies Through Learning. J Artif Soc Soc Simul 9(2) http://www.soc.surrey.ac.uk/JASSS/9/2/9.html
21. Haken H (1978) Synergetics. An Introduction. In: Nonequilibrium Phase Transitions and Self-Organization in Physics, Chemistry and Biology, 2nd edn. Springer, Berlin
22. Harding A (ed) (1996) Microsimulation and Public Policy. Contributions to Economic Analysis, vol 232. Elsevier, Amsterdam
23. Helbing D (1991) A Mathematical Model for the Behavior of Pedestrians. Behav Sci 36:298–310
24. Helbing D (1991/1992) A Mathematical Model for Behavioral Changes by Pair Interactions and Its Relation to Game Theory. Angew Soz 17:179–194
25. Helbing D (1992) A Mathematical Model for Attitude Formation by Pair Interactions. Behav Sci 37:190–214
26. Helbing D (1994) Quantitative Sociodynamics. Stochastic Methods and Models of Social Interaction Processes. Kluwer, Dordrecht
27. Helbing D, Johansson A (2007) Quantitative Agent-Based Modeling of Human Interactions in Space and Time. In: Amblard F (ed) Proceedings of the 4th Conference of the European Social Simulation Association (ESSA'07), Toulouse, 10–14 Sept, 2007, pp 623–637
28. Hutchins E, Hazlehurst B (1995) How to invent a lexicon: the development of shared symbols in interaction. In: Gilbert N, Conte R (eds) Artificial Societies. The Computer Simulation of Social Life. UCL Press, London, pp 157–189
29. Jager W, Popping R, Sande HVD (2001) Clustering and fighting in two-party crowds: Simulating the approach-avoidance conflict. J Artif Soc Soc Simul 4(3) http://www.soc.surrey.ac.uk/JASSS/4/3/7.html
30. Johnson P, Lancaster A (2001) Swarm User Guide. http://www.swarm.org/swarmdocs-2.1.1/userbook/userbook.html
31. Klein C, Fuchs D, Berger P, Hassenpflug P (2006) MicS. Agentenbasierte Mikrosimulation, Project Thesis (Studienarbeit), Universität Koblenz-Landau, Koblenz
32. König A, Möhring M, Troitzsch KG (2003) Agents, Hierarchies and Sustainability. In: Billari F, Fürnkranz-Prskawetz A (eds) Agent-Based Computational Demography. Springer, Berlin, pp 197–210
33. Labrou Y, Finin T (1997) A Proposal for a new KQML Specification, Computer Science and Electrical Engineering Depart-

ment (CSEE). University of Maryland Baltimore County (UMBC), Baltimore

34. Latané B (1981) The psychology of social impact. Amer Psychol 36:343–356
35. Latané B (1996) Dynamic Social Impact. Robust Predictions from Simple Theory. In: Hegselmann R, Mueller U, Troitzsch KG (eds) Modelling and Simulation in the Social Sciences from a Philosophy of Science Point of View. Kluwer, Dordrecht, pp 287–310
36. Latané B, Nowak A (1994) Attitudes as Catastrophes: From Dimensions to Categories with Increasing Involvement. In: Vallacher R, Nowak A (eds) Dynamical Systems in Social Psychology. Academic Press, San Diego, pp 219–250
37. LePage C, Bousquet F, Bakam I, Bah A, Baron C (2000) CORMAS: A multiagent simulation toolkit to model natural and social dynamics at multiple scales. In: Proceedings of the Workshop "The ecology of scales", Wageningen
38. Lewin K (1951) Field Theory in Social Science. Harper, New York
39. Lotka AJ (1925) Elements of Physical Biology. Williams & Wilkins, Baltimore
40. Luke S, Cioffi-Revilla C, Panait L, Sullivan K (2004) MASON: A New Multi-Agent Simulation Toolkit. In: Proceedings of the 2004 SwarmFest Workshop, Santa Fe
41. Lumsden CJ, Wilson EO (1981) Genes, Mind, and Culture. The Coevolutionary Process. Harvard University Press, Cambridge
42. Meadows D et al (1974) Dynamics of Growth in a Finite World. MIT Press, Cambridge
43. Meadows D et al (1992) Beyond the Limits. MIT Press, Cambridge
44. Möhring M (1996) Social Science Multilevel Simulation with MIMOSE. In: Troitzsch KG, Mueller U, Gilbert N, Doran JE (eds) Social Science Microsimulation. Springer, Berlin, pp 123–137
45. North MJ, Howe TR, Collier NT, Vos RJ (2005) The Repast Simphony Runtime System. In: Agent 2005 Conference on Generative Social Processes, Models and Mechanisms. Argonne National Laboratory, Argonne
46. North MJ, Collier NT, Vos JR (2006) Experiences Creating Three Implementations of the Repast Agent Modeling Toolkit. ACM Trans Model Comp Simul 16(1):1–25
47. Orcutt GH, Merz J, Quinke H (1986) Microanalytic simulation models to support social and financial policy. Information Research and Resource Reports, vol 7. North-Holland, Amsterdam
48. Planck M (1930) Positivismus und reale Außenwelt. In: Planck M (1949) Vorträge und Erinnerungen. Hirzel, Stuttgart, pp 228–245
49. Poundstone W (1992) Prisoners' Dilemma. John von Neumann, Game Theory, and the Puzzle of the Bomb. Oxford UP, Oxford
50. Pugh III AL (1976) DYNAMO User's Manual. MIT Press, Cambridge
51. Railsback SF, Lytinen SL, Jackson SK (2006) Agent-based Simulation Platforms: Review and Development Recommendations. Simulation 82(9):609–623
52. Rapoport A (1986) General Systems Theory. Essential Concepts & Applications. Abacus, Tunbridge Wells and Cambridge
53. Richmond B (2001) An introduction to systems thinking: STELLA software. ISEE Systems, Lebanon
54. Rouchier J, Bousqet F, Barreteau O, LePage C, Bonnefoy JL (2000) Multi-Agent Modelling and Renewable Resources Issues: The Relevance of Shared Representations for Interacting Agents. In: Moss S, Davidsson P (eds) Multi-Agent-Based Simulation. Springer, Berlin, pp 181–197
55. Rouchier J, Bousquet F, Requier-Desjardins M, Antona M (2001) A multi-agent model for describing transhumance in North Cameroon: comparison of different rationality to develop a routine. J Econo Dyn Control 25:527–559
56. Sauerbier T (2002) UMDBS – A New Tool for Dynamic Microsimulation. J Artif Soc Soc Simul 5:2–5 http://jasss.soc.surrey.ac.uk/5/2/5.html
57. Schelling TC (1971) Dynamic Models of Segragation. J Math Soc 1:143–186
58. Schelling TC (1978) Micromotives and Macrobehavior. Norton, New York
59. Simon HA (1957) A Formal Theory of Interactions in a Social Group. Amer Soc Rev 17:202–211
60. Simon HA (1957) Models of Man, Social and Rational. Math Essays Ration Hum Behav Soc Setting. Wiley, New York
61. Simon HA (1996) Models of My Life. MIT Press, Cambridge
62. Tobias R, Hofmann C (2004) Evaluation of free Java-libraries for social-scientific agent based simulation. J Artif Soc Soc Simul 7 http://jasss.soc.surrey.ac.uk/7/1/6.html
63. Troitzsch KG (1998) Multilevel Process Modeling in the Social Sciences: Mathematical Analysis and Computer Simulation. In: Liebrand WBG et al (eds) Computer Modeling of Social Processes. Sage, London, pp 20–36
64. Troitzsch KG (2004) Validating Simulation Models. In: Horton G (ed) Networked Simulations and Simulated Networks. 18th European Simulation Multiconference. SCS, Erlangen, San Diego, pp 265–270
65. Volterra V (1926) Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. Allt Accad Naz Lincei 6(2):31–113
66. Weidlich W, Haag G (1983) Concepts and Models of a Quantitative Sociology. The Dynamics of Interacting Populations. Springer, Berlin
67. Wilensky U (1998) NetLogo Segregation model. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston. http://ccl.northwestern.edu/netlogo/models/Segregation
68. Wilensky U (2007) NetLogo 4.0 User Manual. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston. http://ccl.northwestern.edu/netlogo/docs/NetLogo%20User%20Manual.pdf
69. Wooldridge M, Jennings NR (1995) Intelligent agents: theory and practice. Knowl Eng Rev, 10:115–152

Books and Reviews

Ahrweiler P, Gilbert N (1998) Computer Simulations in Science and Technology Studies. Springer, Berlin

Amblard F (2007) Interdisciplinary Approaches to the Simulation of Social Phenomena. In: Proceedings of the 4th Conference of the European Social Simulation Association. IRIT, Toulouse

Casti JL (1997) Would-Be Worlds. How Simulation Is Changing the Frontier of Science. Wiley, New York

Coelho H, Espinasse B (2004) 5th Workshop on Agent-Based Simulation. SCS, Erlangen

Conte R, Hegselmann R, Terna P (eds) (1997) Simulating Social Phenomena. Springer, Berlin

Doran JE, Gilbert N (1994) Simulating Societies: The Computer Simulation of Social Phenomena. UCL Press, London

Edmunds B, Hernández C, Troitzsch KG (2008) Social Simulation. Technologies, Advances, and New Discoveries. Information Science Reference. Hershey, New York

Gilbert N, Conte R (eds) (1995) Artificial Societies. The Computer Simulation of Social Life. UCL Press, London

Gilbert N, Troitzsch KG (2005) Simulation for the Social Scientist. 2nd edn. Open University Press, Maidenhead

Hegselmann R, Mueller U, Troitzsch KG (eds) (1996) Modelling and Simulation in the Social Sciences from a Philosophy of Science Point of View. Kluwer, Dordrecht

Liebrand WBG et al (1998) Computer Modeling of Social Processes. Sage, London

Moss S, Davidsson P (2000) Multi-Agent-Based Simulation. Lecture Notes in Artificial Intelligence, vol 1979. Springer, Berlin

Müller JP, Seidel MM (2003) 4th Workshop on Agent-Based Simulation. SCS, Erlangen

Sichman JS, Conte R, Gilbert N (1998) Multi-Agent Systems and Agent-Based Simulation. Springer, Berlin

Simon HA (1977) Models of Discovery and Other Topics in the Methods of Science. Reidel, Dordrecht

Suleiman R, Troitzsch KG, Gilbert N, Mueller U (1998) Social Science Microsimulation. Tools for Modeling, Parameter Optimization and Sensitivity Analysis. Springer, Heidelberg

Sun R (2006) Cognition and Multi-Agent Interaction. From Cognitive Modeling to Social Simulation. Cambridge UP, New York

Troitzsch KG (2005) Representing Social Reality. In: Pre-Proceedings of the Third Conference of the European Social Simulation Association. Fölbach, Koblenz

Troitzsch KG, Mueller U, Gilbert N, Doran JE (eds) (1996) Social Science Microsimulation. Springer, Berlin

Vallacher R, Nowak A (eds) (1994) Dynamical Systems in Social Psychology. Academic Press, San Diego

Social Psychology, Applications of Complexity to

Robin R. Vallacher
Florida Atlantic University, Boca Raton, USA

Article Outline

Glossary

Dynamical system A dynamical system is a set of interconnected elements that change due to their mutual influences. A change in each element depends on the nature of the influences from other elements. Due to these mutual influences, the system as a whole evolves in time. Because the state of every element is the effect of the states of all the relevant elements in the previous moment, and because the state of every element is one of the causes determining the state of the other elements in the next moment, a dynamical system is characterized by bi-directional causality.

Intrinsic dynamics In a dynamical system, the current state of the system is directly caused by the preceding state of the system. More precisely, the state of each element at a given moment is caused by the states of other elements in the previous moment and by its own previous state. This chain of cause and effect iterated over time gives rise to internally-generated or "intrinsic" dynamics.

Self-organization In a dynamical system consisting of inter-connected elements, the state of each element adjusts to the current state of other elements to which it is connected. Because of this mutual influence, higher order units develop that represent the organization of the basic elements. No higher-order agent is required for such order to emerge. Hence, the process is referred to as self-organization.

Emergence Emergence occurs when the individual elements of a dynamical system achieve organization by means of their mutual influence. The development of the higher level state is said to be emergent because this state was not inherent in the properties of the lower-level elements and because the higher-level state was not imposed on the system from forces outside the system. The higher-order properties that result from the mutual adjustment among lower-level elements provide coordination for the lower-level elements. Emergence thus provides for substantial growth in the complexity of a system's processes and properties. Because of emergence, very complex systems can often be described by very simple models.

Attractor An attractor represents a subset of a system's phase space to which the system evolves over time and which resists forces that would perturb this temporal trajectory. This subset can consist of a region of nearby states or it can represent two (or more) regions among which the system oscillates over time in a periodic, quasi-periodic, or chaotic manner. In a system governed by attractor dynamics, a relatively wide range of starting points (initial states) will eventually converge on a much smaller set of states or on a pattern of change between states defining the attractor. External influences may push the system out of the attractor, but over time the system will return to this equilibrium.

Cellular automata Cellular automata are programmable dynamical systems. In this approach, a finite set of elements is specified, each of which can adopt a finite number of discrete states. The elements are arranged in a specific spatial configuration that usually takes the form of a two-dimensional lattice or grid. The location of each element on this grid specifies the element's neighborhood. The elements evolve in discrete units of time, such that the state of an element at $t + 1$ depends on the states of the neighboring elements at time t. The dynamics of cellular automata depend on the nature of the updating rule and on the format of the grid dictating the neighborhood structure.

Definition of the Subject

The seminal insights that launched the field of social psychology over a century ago have stood the test of time. Such early scholars as James [41], Cooley [21], Mead [65], Lewin [59], and Asch [4] emphasized the multiplicity of interacting forces operating in individual minds and in social groups and the potential for sustained patterns of change resulting from such complexity. The inherent complexity and dynamism of human experience, however, proved problematic for mainstream social psychology during much of the 20th century. The canonical paradigm in this period reduced dynamics to a one-step process involving a purported cause, operationalized as an independent variable at an arbitrary time 1, and its effect, operationalized as a dependent variable that was assessed at an arbitrary time 2. The complexity of experience, with multiple variables interacting over time to generate a stream of thought or action, was difficult to investigate with available statistical techniques (e. g., correlation, regression, analysis of variance). Social psychology thus fostered an image of human experience that emphasized simplicity and stability at the expense of complexity and dynamism.

Contemporary social psychology shows signs of returning to the deep intuitions concerning human experience articulated by the field's founding fathers. The renewed appreciation for complexity and dynamism was made possible by developments in the understanding of nonlinear dynamical systems in the 1970s and 1980s, and the application of these developments to social processes within the last two decades. Computer simulations capture the complexity of social processes and document the emergence of higher-order properties from the interaction of basic elements in a mental or social system. Innovative means of collecting and analyzing time-series data provide rigorous insight into the intrinsic dynamics of mental, affective, behavioral, and interpersonal processes. And formal models enable researchers to identify the parameters that are critical for understanding the dynamism and complexity of social psychological phenomena. In short, the contemporary emphasis on complexity and dynamics provides a new paradigm for the field, one that holds promise for advancing the social psychology as a precise science while preserving the basic insights that launched the field over a century ago.

Introduction

Despite its ubiquity, dynamism is a poor candidate for theory construction. Theories are expressed in terms of invariant properties representing stable "signals" that are obscured by the "noise" associated with personal, interpersonal, and societal processes. The stream of consciousness may be an accurate depiction of subjective experience, for example, but its turbulent nature seems inconsistent with fundamental properties that transcend particular individuals and their moment-to-moment mental states. To identify regularities and invariant properties, social psychologists commonly ignore the ever-changing undercurrent of mind and action, focusing instead on those elements of thought and behavior that admit to stability and structure. Accordingly, they typically emphasize higher order units of mental and behavioral phenomena (e. g., traits, schemata, global evaluations, norms) that presumably lend stability and coherence to subjective experience. The turbulent flow of lower level elements (thoughts, feelings, movements) is effectively rendered irrelevant to "true" understanding and meaningful prediction.

The gulf between the reality of experiential turbulence and the focus on stability in theory construction is unnecessary. Dynamism and structure represent complementary aspects of experience that together provide the basis for cognitive, emotional, and behavioral accommodation. The basic notion is that the flow of lower-level elements gives rise to higher order structures, which in turn constrain the dynamics of lower-level elements. This reciprocal causal relationship between lower-level and higher-level units of experience is central to *dynamical social psychology* [74,106] a recently developed paradigm that represents an adaptation of the concepts, principles, and tools developed within dynamical systems and complexity science.

It is not feasible to describe every version of the dynamical perspective, nor every application of this general approach. This may have been possible in the

1990s [26,31,35,45,102] when the dynamical perspective was a promissory note rather than an established paradigm. Since that time, there has been a proliferation of many innovative research strategies—some providing formal models implemented in computer simulations, others offering empirical means for investigating the dynamics of personal, interpersonal, and societal processes—that have been used to investigate a wide variety of phenomena. The aim herein is to highlight the crucial elements of the dynamical perspective that find expression in otherwise distinct theories, research strategies, and topical agendas.

The first two sections (Sects. "Intrinsic Dynamics" and "Attractors in Psychological Systems") describe basic concepts from the study of nonlinear dynamical systems that are directly relevant to social psychology. Section "Dynamical Minimalism" describes dynamical minimalism [72], an approach that provides a workable entrée into the nature and expression of dynamic processes at different levels of social reality. This approach is illustrated in the next two sections with respect to two research agendas—one emphasizing the emergence of group-level properties from the self-organization of individual agents (Sect. "The Dynamics of Social Influence") the other exploring the tendency of individuals to coordinate their behavioral and mental dynamics in service of forming dyads and social groups (Sect. "Dynamics of Interpersonal Coordination"). The final section ("Future Directions") reflects on the trajectory of dynamical social psychology thus far and offers caveats concerning the relevance of this approach to the unique features of human experience.

Intrinsic Dynamics

People's mental and emotional states, overt behavior, and social relations evolve and change in the absence of external influence. The ubiquity of intrinsic dynamics is apparent at different levels of social reality, from basic intrapersonal processes to macro-level societal phenomena. At the level of the mind, the temporal pattern of cognitive and affective elements in the stream of thought [41] often provides a more accurate depiction of a person's mental make-up than do the summary aspects of the person's mental process (e. g., overall attitude, final decision) that are more often the focus of investigation [103]. For example, research has shown that simply thinking about an attitude object (e. g., another person) in the absence of external influence or new information tends to promote more extreme (polarized) evaluations of the object over time [97].

Research on social judgment has shown that internally generated thoughts and feelings about a target person often reflect elaborate but identifiable patterns of change that convey important information. A judgment that is neutral when average over time, for instance, can have very different meanings and implications, depending on the intrinsic dynamics of the judgment process [109]. When neutrality represents little variation in evaluation occurring on a relatively slow time-scale, the summary judgment might well reflect a truly neutral sentiment. If neutrality reflects oscillation between highly positive and highly negative judgments on a rapid timescale, however, the summary judgment signifies heightened involvement and ambivalence rather than neutrality per se.

Intrinsic dynamics also characterize personal action. Actions have a hierarchical structure, in that the performance of an action entails the coordinated interplay of more basic actions or sub-acts. "Going to work," for example, may involve getting dressed, leaving the house, driving a car, parking the car, and entering a building. Each of the lower-level acts, in turn, can be decomposed into yet more basic lower-level elements. "Driving," for instance, consists of starting the car, turning the steering wheel, making turns, and braking. Each level in an action hierarchy is associated with a different time scale, with increasingly lower-level acts taking place in correspondingly shorter intervals of time [69]. "Going to work" unfolds on a longer time scale than does "driving," for example, and the time scale for "driving" is longer than that for each instance of "turning the steering wheel." There is evidence that the embedded time scales in an action hierarchy often have a fractal structure [69].

The intrinsic dynamics of action, in turn, span the levels of action in an overall action hierarchy. What may appear to be a continual succession of momentary movements when defined in low-level, mechanistic terms can take on the appearance of switching between qualitatively different actions, each occurring on a longer time scale, when defined in higher-level terms. Research on action identification [107] has identified several factors that determine the level at which an action is regulated. This line of work has also demonstrated that people reliably differ in their default level of action identification across many action domains [108]. Individuals who tend to think about their actions in lower-level terms are predisposed to the emergence of higher-level action understanding. Emergence occurs when such individuals are exposed to cues (e. g., feedback from other people) that suggest higher-level meaning for the actions [114] or when they privately reflect on their actions, allowing the lower-level action elements to self-organize into a higher-level act identity that provides subjective integration for these elements [104].

Social interaction has also been investigated with respect to intrinsic dynamics. For example, research has focused on the interpersonal coordination of relatively low-level actions, such as speaking (e. g. [20,23]) and limb movement (e. g. [6,48,69,101]). In one approach, two individuals are asked to swing their legs while sitting down across from one another [6]. One person swings his or her legs in time to a metronome and the other person tries to match those movements. This research has revealed two forms of coordination: *in-phase*, with the individuals swinging their legs in unison, and *anti-phase*, with the individuals swinging their legs with the same frequency but in the opposite direction. Individuals can maintain anti-phase coordination only up to a certain frequency of movement, at which point they switch to in-phase coordination. When the frequency is subsequently decreased, at some value they are able to coordinate anti-phase again, but this tempo is significantly lower than the point at which they originally started to coordinate in-phase. Such *hysteresis* indicates that movement coordination can be analyzed as a nonlinear dynamical system [38,48]. This line of research has identified modes of coordination more complex than in-phase and anti-phase [5,87,101].

Interpersonal dynamics are not confined to the coordination of speech and motor movements, but also include the temporal coordination of higher-level actions (e. g., plans, goals) and internal states (e. g., moods, judgments). Although this topic has not been heavily researched, there is evidence that the quality of a social relationship is reflected in partners' ability to coordinate in-phase with respect to their respective higher-level actions, opinions, and feelings (e. g. [5,36,64,82,83,100]). The ebb and flow of feelings, information exchange, and action conveys deeper insight into the nature of a relationship than do global indices such as the average sentiment, the amount of information exchanged, or the summary action tendencies. Colloquially, people who feel positively about one another are said to "be in synch" or "on the same wavelength" with respect to their internal states.

At the societal level, tracking the temporal trajectory associated with the emergence of norms and public opinion provides greater insight into the society's future make-up and likely response to external threat than simply knowing what the societal norms and opinions are at a single point in time [76]. When norms and opinions develop gradually over a long period of time, for example, the society displays resistance to external threats or even to new information that might promote better economic conditions. But societal change in political and economic ideology can occur in a rapid, nonlinear manner (e. g. [75,84]), with a temporal trajectory that resembles phase transitions in physical systems [58]. Societies that undergo such nonlinear transitions are vulnerable to later rebounds of the earlier ideologies and are highly responsive to threats and new information, and they can experience a period of sustained oscillation between conflicting worldviews [75].

Attractors in Psychological Systems

Psychological systems are characterized by intrinsic dynamics, but they also demonstrate stability and resistance to change. Each day, people encounter vast amounts of information relevant to social judgment and interpersonal relations, much of which is mutually contradictory. People nonetheless manage to form and maintain coherent patterns of thought and behavior in their social lives. The partners to a romantic relationship, for example, experience a wide variety of thoughts and feelings about one another, but over time each partner's mental state tends to converge on positive sentiment toward the other. Thus, despite the ever-changing nature of intrapersonal and interpersonal experience, people's mental, emotional, and behavioral states tend to converge on relatively narrow sets of specific states or on patterns of change between specific states. These states or patterns of change are referred to as *attractors*.

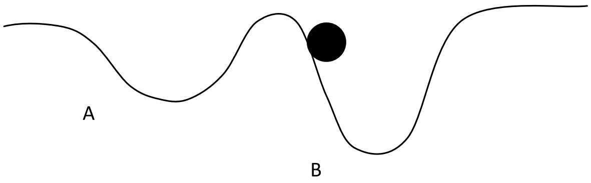
When a system is at its attractor, it tends to maintain that state despite potentially destabilizing forces and influences. An external influence may move the system to another state, but the system will return fairly quickly to one of its attractors. Several social psychological phenomena imply the existence of an attractor. For example, self-regulation is defined in terms of resistance to temptations and distractions, impulse control, and the maintenance of states representing salient personal standards and values [14,15,105]. In the same fashion, self-esteem maintenance [98], self-verification [95], and psychological reactance [8] all reflect the tendency of mental systems to converge on a particular state (e. g., a level of self-esteem) and to resist influences that threaten to dislodge the person's judgments and beliefs from that state.

Three basic types of attractors are commonly distinguished: fixed-point attractors, periodic (including multiperiodic) attractors, and deterministic chaos (cf. [25,73,89]). Each type is likely to have relevance for understanding different intrapersonal and interpersonal processes, although work designed to establish such relevance is in its nascent stage. A system can be characterized by another set of points that have the opposite effect of attractors. Termed *repellors*, they represent unstable equilibria of the system, so that the smallest departure from their exact values will result in the system rapidly escaping from the region surrounding that value. Whereas attractors may be described as states that the system "seeks" over long periods of time, repellors reflect states that the system selectively "avoids." The repellor concept has been used to describe how neural networks avoid particular memories [105].

Fixed-Point Attractors

A *fixed-point attractor* is similar to the notion of equilibrium or homeostasis [12,66]. It represents the case in which the state of the system converges to a stable value. A desired end-state or goal [14,15,104,106] for example, can be described as a fixed-point attractor. This is apparent when a person maintains a belief, an interpersonal evaluation, or an action tendency despite forces or sources of information that challenge these tendencies. Attractors, however, are not limited to goals, intentions, or other desired states. Thus, a person might display a pattern of antagonistic behavior in his or her social relations, despite efforts to avoid behaving in this manner. Similarly, a person with low self-esteem may initially embrace flattering feedback from someone, but over time the person may discount or reinterpret this feedback, displaying instead a pattern of self-evaluative thought that converges on a negative state [96]. In an inter-group context, meanwhile, warring factions may display conciliatory gestures when prompted to do so, but revert to a pattern of antagonistic thought and behavior when the outside interventions are relaxed [19]. In short, when a system's dynamics are governed by a fixed-point attractor, the system will consistently evolve to a particular state, even if this state is not hedonically pleasant, and will return to this state despite being perturbed by forces that might promote a more pleasant state.

A psychological system may have more than one attractor, with each corresponding to a distinct equilibrium. Which attractor governs a system's dynamics in a particular instance depends on the starting values of the system's evolution. The set of initial states that converge on each attractor represents the *basin of attraction* for that attractor. For a person or a group characterized by multiple fixed-point attractors, then, the process in question can display different equilibrium tendencies, each associated with a distinct basin of attraction. Within each basin, different initial states will eventually converge on the same stable value. However, even a slight change in the system's initial state will promote a large change in the system's trajectory if this change represents a state that falls just outside the original basin of attraction and within a basin for a different attractor. For example, in conflict situations there are typically two dominant responses, one corresponding to aggression and one corresponding to conciliation. Slight differences in the circumstances surrounding the conflict can thus promote dramatically different behaviors, with no option for a response that integrates the two tendencies (e. g. [19]).



A dynamical system with two fixed-point attractors (A and B)

Social Psychology, Applications of Complexity to, Figure 1
A dynamical system with two fixed-point attractors [106]

The attractor concept can be captured in a simple metaphor. Figure 1 shows a ball on a hilly landscape. The ball represents the current state of the system and the valleys (A and B) represent different fixed-point attractors. The evolution of the system toward an attractor is characterized by the ball rolling down a hill and coming to rest in the bottom of a valley. Each attractor in Fig. 1 has a basin of attraction, represented by the width of the valley. The basin of attraction for Attractor A in Fig. 1 is wider than the basin for Attractor B, which means that a wider range of states will evolve toward Attractor A. Attractors can also vary in their respective strength, which is represented by the corresponding depth of the two valleys. Attractor B, then, is stronger than Attractor A. Hence, it is more difficult for external influence to change the current state of the system when the system is within the basin of attraction for Attractor B as opposed to Attractor A.

The existence of multiple fixed-point attractors in a system captures the intuition that people may have different (even mutually contradictory) goals, self-concepts, recurring emotional states, and patterns of social behavior. Someone may have two or more standards for self-regulation, for example, with each providing for action guidance under different sets of conditions. The person's behavior may reflect an affiliation standard under one set of conditions, but reflect an achievement standard under a different set of conditions. In similar fashion, a person may have multiple self-views (e. g. [37,60,61,93]), each representing an integrated and stable way of thinking about him or herself. One self-view is likely to become salient when a specific set of self-relevant information is made salient by virtue of context or role expectations. Characteristic patterns of emotional response, meanwhile, can be viewed as distinct attractors, with each attractor reflecting an "emotional Gestalt" that provides coherence for a person's cognitive-affective dynamics under a unique set of conditions and constraints [57,99]. More generally, apparent inconsistency in personality can be viewed as the existence of multiple attractors, each associated with a different basin of attraction for thought, feeling, and action [82,86,92]. One set of conditions might promote a trajectory that evolves toward dominance and competition, for instance, whereas another might promote empathy and cooperation.

The strength of an attractor and the size of its basin of attraction are potentially independent, and different combinations of these basic properties may have unique implications for psychological processes [82,83]. Note again the two attractors in Fig. 1. Attractor A is relatively weak but has a wide basin of attraction. Thus, a relatively small force may change the state of the system (i. e., move the ball up the gradual slope), but the system is likely to return to the attractor (i. e., it will roll back into the valley) even if these changes are relatively large. In contrast, Attractor B has a narrower basin of attraction but is relatively strong. A great deal of force is required to promote even a slight impact on the system (i. e., move the ball up the steep slope), but if this effect is achieved, the system will not return to the attractor (i. e., it will escape the valley).

As an example, consider a romantic couple that has two attractors: a strong attractor associated with positive feelings and a weak attractor associated with negative feelings. Assume the couple has a wider range of attraction for positive feelings than for negative feelings. The partners are likely to evolve toward positive feelings about one another if they begin an interaction within a broad range of affective states (e. g., neutral to very positive), but they may end up feeling negative if they begin an interaction within a different (more restricted) range of affective states (e. g., mildly to highly negative). A broader range of initial states are likely to promote a communication trajectory that results in an exchange of warm sentiments as opposed to critical comments. If, however, the couple typically starts out with negative feelings, the negative attractor, despite having a narrow basin, may dictate the trajectory for feelings expressed in the couple's interactions. It is conceivable, of course, that the couple has a wider basin of attraction for negative feelings, so that anything short of a highly positive initial state will dissolve into a negatively toned exchange (see [33,34]).

Latent Attractors

A system may have multiple attractors, but when the system is at one of them, the others may not be visible to observers, or perhaps not even to the actors themselves. The existence of these potential states might not even be suspected. Such *latent attractors* may be highly important in the long run, though, because they define which states are possible for the system when conditions change. By identifying possibilities for a system that have yet to be experienced, the concept of latent attractor goes beyond the traditional notion of equilibrium (e. g. [2]). Critical changes in a system might not be reflected in the manifest state of the system, but rather in the creation or destruction of a latent attractor representing a potential state that is currently invisible.

The implications of latent attractors have been explored in the context of inter-group characterized by seemingly intractable conflict [18,19,85]. Factors such as objectification, dehumanization, and stereotyping of outgroup members are preconditions for the development of intractable conflict [17,22] but their immediate impact may not be apparent. Rather, these factors may create a latent attractor to which the system can abruptly switch in response to a provocation that is relatively minor, even trivial. By the same token, efforts at conflict resolution that seem fruitless in the short run may have the effect of creating a latent positive attractor for inter-group relations, thus establishing a potential relationship to which the groups can switch if other conditions permit. In this case, the existence of a latent positive attractor might promote a rapid de-escalation of conflict, even between groups with a long history of and seemingly intractable conflict.

Periodic Attractors

Rather than converging on a stable value, systems can display sustained rhythmic or oscillatory behavior. Such a temporal pattern is referred to as a periodic or limit-cycle attractor. Periodicity is obviously associated with many biological phenomena, such as circadian rhythms and menstrual cycles [30], but this dynamic tendency may also underlie certain psychological phenomena [32]. Moods have been shown to have a periodic structure, for example, often corresponding to a weekly cycle (e. g. [9,50,51]). Research on the intrinsic dynamics of both social judgment [109] and self-evaluation [110], meanwhile, has demonstrated that the stream of thought often oscillates between positive and negative assessments, sometimes in accordance with remarkably fast time scales. Periodic structure also characterizes human action [69] and is a feature of social interaction as well (e. g. [6,32,70]).

Distinguishing a periodic attractor from the existence of multiple fixed-point attractors can prove difficult, since the system in both cases displays movement between different states over time. The difference centers on the regularity of the movement between states and the role of external factors in producing such movement. A periodic attractor reflects a repetitive temporal pattern, such that the values of the dynamical variable repeat after a time $T, x_i(t) = (t + T)$, where T is the period of motion. The state of the system undergoes constant change even in the absence of noise or external influence. To qualify as a periodic attractor, then, a pattern of change must represent a pattern on which the system converges, and to which it returns after small perturbations. In a daily activity cycle, for example, a sleepless night might temporarily disrupt the pattern (e. g., oversleeping the next few days), but eventually the pattern will be restored.

A system with fixed-point attractors, in contrast, tends to stabilize on a particular state or set of states. Such attractors capture all trajectories within their respective basins, so a disturbance, noise, or an external influence is necessary to move the system from one stable state to another. A person with self-regulatory standards for both compromise and confrontation, for example, will display one of these tendencies as long as the action context is within the basin of attraction for that tendency. If the attractors differ in the size of their respective basins, and if contexts are avoided that attract the person's mental, emotional, or behavioral state toward the smaller basin, the person may behave for long periods of time in line with the stronger attractor. In similar fashion, a romantic couple may have fixed-point attractors for both positive and negative affective states, but whether they display periodic movement between them will depend on the starting conditions associated with their interactions. Even if the couple oscillates between positive and negative states, each state provides at least temporary stability. In periodic evolution, stability is not provided by any particular state, but rather by the pattern of changes between states.

The distinction between fixed-point and periodic attractors was observed in a study investigating the temporal trajectories of affective states on the part of bipolar depressive individuals [42]. Time-series analysis of mood and other symptoms revealed that many of these patients oscillated between a normal and a depressed state. However, patients whose temporal dynamics did not reflect fixed-point attractor tendencies were at highest risk for suicide and were hospitalized more often for their depression. These risks were low for individuals whose moods oscillated around a single attractor, even one corresponding to a depressed state, and for those whose moods switched between two distinct attractors reflecting a normal state and a depressed state. These results imply an interesting connection between attractor dynamics and self-regulatory tendencies. Self-regulation implies stabilization with respect to some states and de-stabilization of other states. The stable states reflect fixed-point attractors for a person's mental and emotional dynamics. The lack of fixed-point attractors for one's internal state signals a breakdown in the capacity for self-regulation.

Deterministic Chaos

The best known phenomena concerning nonlinear dynamical systems centers on *deterministic chaos* (cf. [89]). Many researchers and scholars, especially those from fields other than mathematics and physics, commonly discuss the primary insights from the work on nonlinear dynamics as chaos theory. When investigating a chaotic system, anything less than infinite precision in the knowledge of a system at one point in time can undermine prediction of the system's future states. This decoupling between determinism and practical predictability happens because all initial inaccuracies are amplified by the system's intrinsic dynamics, so that the inaccuracies grow exponentially over time. After some time, exponential growth assures that the size of the error will exceed the possible range of states of the system's behavior.

Chaos clearly represents a possibility in nonlinear dynamical systems [31,73], and has been demonstrated in many biological and physical phenomena. In principle, then, human thought and behavior may sometimes follow a chaotic trajectory. Despite this potential, though, unequivocal evidence for deterministic chaos in human thought and behavior remains to be documented. Human dynamics always contain some degree of randomness and human behavior is often unpredictable. It can be quite difficult, however, to determine the degree to which such unpredictability reflects deterministic chaos, the stochastic nature of the laws governing human nature, or the multitude of influences unaccounted for by measurement that can be treated as noise.

Dynamical Minimalism

In canonical social psychology, the complexity of human thought and behavior is typically assumed to reflect complex interactions among a large number of variables in the traditional approach to theory construction. The approach of *dynamical minimalism* [72], in contrast, assumes that complex properties can result from simple rules specifying the interactions among very simple elements. The emergence of complexity occurs when the elements are nonlinear and interact over time [39]. For example, complex cognitive phenomena (e. g., pattern recognition, error correction), have been observed in a simple network of binary elements, where each element reacts to the input it receives from other elements [40]. In this spirit, dynamical minimalism attempts to identify the simplest set of assumptions capable of producing a phenomenon of interest. In formal models, this is equivalent to identifying the simplest mathematical rules to express what is known about a phenomenon. The goal of dynamical minimalism, then, is to achieve parsimony in theory construction without stripping the phenomenon of its subtlety and nuance.

This approach provides a new perspective on the relation between micro and macro levels of description. Reductionism is commonly assumed in mainstream models, such that the rules observed at one level of description correspond to the rules operating at another level. In effect, the properties at a macro level of description are reduced to the properties of elements at a micro level. One might, for example, explain the relation between poverty and crime in a social system by reducing this relation to the relation between frustration and crime at the level of individuals in the society. Dynamical models, however, do not assume isomorphism among levels of description. To the contrary, the rules specifying the interaction among a system's elements are likely to generate very different rules at higher levels of system behavior (see [24] for an early appreciation of this idea).

The emergence of new properties at a macro level is illustrated in the *society of self* model of self-structure [81]. This model assumes very simple rules by which self-relevant information is integrated in forming a self-concept. Each element of information (e. g., an episodic memory, a physical feature, a self-perceived trait) adopts the prevailing valence (positive vs. negative) of related elements. This simple rule, when iterated over time, generated interesting but largely unexpected consequences at the global level of self-understanding. The self-structure became differentiated into locally coherent regions (e. g., social roles, areas of competence), each of which displayed resistance to discrepant information (e. g., negative social feedback). Global self-esteem and high self-concept certainty also emerged at the macro level from the simple rule of influence among elements at the micro level.

Emergence seems to represent a paradox for theory construction. How can knowledge of the lower-level elements provide an explanation of the higher-level properties if properties at a macro level cannot be derived from properties of the system's lower level elements? The role of computer simulations in dynamical minimalism helps to resolve this paradox. With computer models, one can specify the properties of system elements and the rules of interaction among these elements. As the elements interact in accordance with these rules over time, dynamics appear at the system level that were not assumed or programmed for the elements themselves. With computer simulation, then, a theory constructed at a basic level of psychological reality (e. g., moment-to-moment thought process, dyadic social interaction) can be tested at a higher level of psychological reality (e. g., social judgment, group norms).

Computer simulations play another important role in dynamical minimalism. The basic elements comprising a system are often uninteresting and trivial, and the interactions among them may have only minor impact on the systems' global properties. But some properties of system elements may have an important impact on the system's higher-order properties as the elements interact over time. It may not be obvious which properties are trivial and which are essential for the emergence process. Dynamical minimalism makes this distinction and thus constructs a model that incorporates only the properties that are critical for the emergence of macro level phenomena. Computer simulations enable one to systematically vary the assumptions regarding the properties of elements and their interactions, and then observe which assumptions promote meaningful changes at the macro level. Those properties that have trivial consequences at the macro level are eliminated from the model. In short, computer simulations enable one to distill the minimal set of components necessary to capture the essence of a phenomenon.

Computer simulations cannot substitute for empirical verification of a theoretical model. They are critical in identifying the properties that are central to the model and investigating the consequences of these properties for the functioning of the system in question. Knowledge of these consequences then provides the basis for framing hypotheses to be investigated in empirical research. The relationship between computer simulations and empirical research work in the other direction as well. Empirical studies can refine a model, which is then implemented in computer simulations. The results of the simulations, in turn, may generate new hypotheses to be tested in subsequent empirical research. The reciprocal feedback loops among theory, computer simulation, and empirical research is central to the approach of dynamical minimalism.

The Dynamics of Social Influence

Social influence refers to any change in an individual's thoughts, feelings, or behavior that occurs as the result of the real or imagined presence of others [3]. It is widely considered to be the core process of social experience [111] and is manifest in a wide array of phenomena, including obedience to authority, conformity, imitation and modeling, bystander intervention in emergencies, social loafing, stage fright, persuasion, and groupthink. There is evidence, however, that the essence of social influence reflects the interplay of three basic factors: the number, strength, and immediacy of the sources of influence [52]. Empirical research has shown that the magnitude and nature of social influence represents a multiplicative function of these sources.

Nowak et al. [76] modeled the dynamics of social influence using cellular automata. In this approach, each individual is represented with three properties: an opinion on a topic, a degree of persuasive strength, and a position in a social space. Individuals are commonly assumed to have one of two opinions on an issue (e. g., pro vs. con). The group consists of n individuals located on a two-dimensional grid, with each cell corresponding to an individual (see Fig. 2). The color of each cell specifies that individual's current opinion (light gray denotes pro, dark gray denotes con), whereas the height of the cell represents the individual's strength (e. g., expertise, confidence, charisma). Each individual discusses the issue with other group members ascertains the degree of support for each position. As a result of these assessments, each individual adopts the opinion that is most prevalent. The following formula expresses the strength of influence of each opinion:

$$I_i = \left(\sum_1^N \left(\frac{s_j}{d_{ij}^2} \right)^2 \right)^{1/2} ,$$

where I_i denotes total influence, s_j represents the strength of each individual, and d_{ij} represents the distance between individuals i and j. The opinions of those who are closest to the individual and have the greatest strength are weighted most heavily by the individual. An individual's own position is also considered and is weighted most heavily because of its immediacy (0 distance). Influence grows with the square root of the number of people exerting influence.

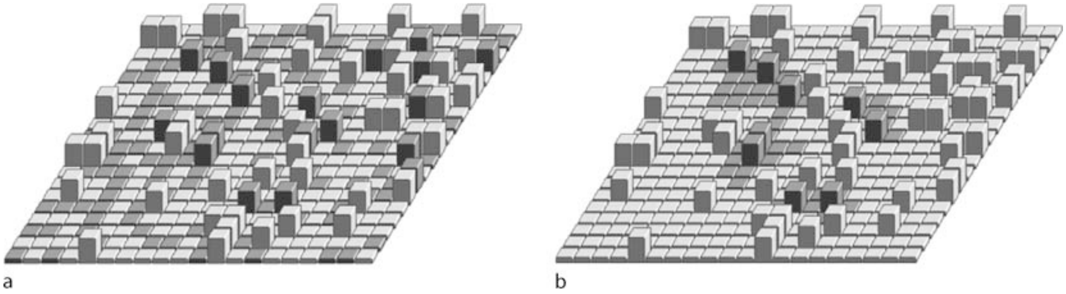
In each round of the simulations, one individual is chosen at random and influence is computed for each opinion in the group. A simple updating rule is employed: the individual changes his or her opinion to match the prevailing opinion if the resultant strength of this opinion position is greater than the strength of the individual's current position. This basic process is performed for each individual in the group. This procedure is repeated until there are no further changes in opinion. This typically involves several rounds of simulation, because an individual who had previously changed his or her position to match that of his or her neighbors may revert to the original position if the neighbors change their opinions.

Figure 2 depicts representative results of the computer simulations. In the initial configuration (Fig. 2a), there is a majority of 60% (light gray) and a minority of 40% (dark gray), with the majority and minority opinions randomly distributed in social space. The majority and minority groups have the same relative proportions of strong and weak members (tall vs. short cells). Figure 2b shows that an equilibrium is reached after six rounds of simulated discussion. Although the majority has grown (to 90%) at the expense of the minority (now 10%), the minority opinion has survived by forming clusters of like-minded people. Note that these clusters are primarily formed around strong individuals.

These two group-level outcomes—polarization and clustering—are routinely observed in computer simulations of this process [55]. Each is reminiscent of well-documented social processes. Research on group dynamics (e. g. [68]), for example, has shown that the average attitude in a group becomes polarized in the direction of the prevailing attitude as a result of group discussion. In the model, polarization reflects the greater influence of the majority opinion. In the initial (random) configuration (Fig. 2a) the average proportion of neighbors holding a particular opinion (pro or con) reflects the proportion of this opinion in the total group. The average group member, in other words, is surrounded by more majority than minority members, so more minority members are converted to the majority position than vice versa. However, some majority members are converted to the minority position because they happen to be located close to an especially strong minority member or because more minority members happen to be at this location.

Clustering is also a pervasive feature of social life, having been documented for such diverse facets of social life as attitudes, political beliefs, religions, clothing fashions, and farming techniques. It has been show, for example, that attitudes tend to cluster in residential neighborhoods [28]. When opinions are distributed randomly, the sampling of opinions through social interaction provides a realistic portrayal of the distribution of opinions in the larger society. However, when opinions are clustered, the same sampling process produces a very biased result because the opinions of one's nearby neighbors are weighted the most heavily. The prevalence of one's own opinion is therefore likely to be over-estimated. Opinions that are in the minority in global terms, then, can form a local majority. This process enables individuals with a minority opinion to maintain this opinion in the belief that it actually represents a majority position.

In the spirit of dynamical minimalism, this model focuses on the minimal set of processes responsible for the emergence of group-level properties that are invariant across diverse areas of topical interest. In so doing, it pro-



Social Psychology, Applications of Complexity to, Figure 2

a Initial distribution of opinions in the simulated group. b Final equilibrium of opinions in the simulated group [106]

vides a platform for investigating how group-level properties emerge for different domains of social functioning. This potential has been developed in recent years. Kenrick and his colleagues, for example, have simulated the emergence of cultural norms from the mutual influence among individuals with conflicting strategies (decision-rules) pertaining to fundamental human goals (e. g. [49]). These goals, reflecting domains of adaptive functioning confronted by human groups throughout history [10,11,29], include self-protection, coalition formation, status-seeking, mate choice, relationship maintenance, and offspring care.

In their computer simulations, individuals with different decision-rules regarding a particular domain interacted with one another over time, and the group-level consequences of these interactions were observed. One series of simulations investigated how individual differences in decision-rules for cooperation versus competition affected community level propensities for such behavior. Another series explored the mutual impact of male and female mating strategies on the emergence of societal-level norms regarding mating. Clustering was observed in both cases, with an initial random configuration of decision rules giving way over time to local communities characterized by coherent norms regarding the behavior in question. In the mating simulations, for example, some communities were characterized by relatively unrestricted mating strategies (reflecting the decision-rule preferred by males), whereas other communities developed norms sanctioning restricted mating strategies (reflecting the decision-rule preferred by females).

Dynamics of Interpersonal Coordination

The dynamical account of social influence provides a concise description of how the state (e. g., attitude) of a single individual depends on the state of other individuals. Because many psychological processes reflect intrinsic dynamics, however, individuals can be conceptualized as displaying patterns of change rather than as a set of states. In this view, social influence (and social interaction generally) can be investigated as the coordination over time of individual dynamics.

The most basic form of coordination is positive correlation or in-phase relation. In social interaction, this occurs when the overt behaviors, attitudes, or emotions of one person induce similar behaviors or states in the other person at the same time (e. g. [16]). This basic form of coordination is reflected in such familiar phenomena as imitation, mimicry, and empathy [62]. Other forms of coordination can be identified, however, with counterparts in different contexts for social interaction [62,69]. Turn taking in conversation represents negative correlation or anti-phase relation between individuals in their respective talking and listening (when one person speaks, the other is silent). Negative correlation also characterizes antagonistic relationships, in that the sadness or despair of one person induces satisfaction or happiness in the other person and vice versa. More complex forms of synchronization can also be identified that reflect nonlinear relationships and higher-order interactions between the partners' respective behaviors and internal states (cf. [74,82]).

A Model of Synchronization Dynamics

This perspective provides the foundation for a recently developed model of *synchronization* [82,83], a phenomenon that characterizes coupled dynamical systems [44,91]. Because positive correlation represents the most fundamental and common form of coordination, it has provided the primary focus to date in this class of models. The Nowak et al. [82,83] model assumes that each individual attempts to achieve synchronization by adjusting his or her internal state or overt behavior in response to the state or behavior of the individual with whom he or she is interacting. Individuals in social interaction, in other words, modify their respective thoughts, feelings, or action tendencies to promote positive correlation over time in these features of experience. The synchronization of individuals' dynamics results in a higher order system with its own dynamic properties.

Coupled logistic are used to model interpersonal synchronization [74,81,82,83]. The dynamics of each individual are represented with a logistic equation, which is the simplest dynamical system capable of displaying complex (e. g., chaotic) behavior [27,89]. However, the behavior of each person not only depends on his or her preceding state but also to a certain extent on the preceding state of the other person. The coupling of individuals' dynamics is specified in the following equation:

$$x_1(t+1) = \frac{r_1 x_1(t)(1-x_1(t)) + \alpha r_2 x_2(t)(1-x_2(t))}{1+\alpha}$$
$$x_2(t+1) = \frac{r_2 x_2(t)(1-x_2(t)) + \alpha r_1 x_1(t)(1-x_1(t))}{1+\alpha} .$$

The dynamical variable (x) represents the intensity of behavior, and the control parameter, r, corresponds to internal states (e. g., personality traits, moods, values) that shape the person's pattern of behavior (i. e., changes in x over time). To the value of the dynamical variable representing one individual's behavior x_1, one adds a fraction α of the value of the dynamical variable representing the other individual's behavior x_2. The magnitude of α represents the strength of coupling and can be viewed as the degree of mutual influence characterizing the interaction. Depending on the social context, it might reflect the intensity of communication or the degree of mutual imitation. When α is 0, there is no coupling (e. g., no influence or communication) on the behavior level, whereas when α is 1, each individual's behavior is determined equally by his or her preceding behavior and the influence of the other individual. Intermediate values of α represent moderate values of coupling in the relationship.

Modeling Behavioral Synchronization

The respective control parameters of two individuals are unlikely to be identical when they first interact with one another. And the degree of influence between individuals differs across interactions and relationships. To represent this variability in social reality, Nowak, Vallacher [74] systematically varied the similarity of partners' control parameters (r), representing their internal states, and their degree of coupling (α), representing their mutual influence (e. g., communication, imitation). Each simulation began with a random value of x for each individual, drawn from a uniform distribution that varied from 0 to 1. They let the simulations run for 300 steps, allowing each system to converge on its pattern of intrinsic dynamics and both systems to synchronize. For the next 500 simulation steps, they recorded the values of x for each system and the degree of synchronization of the two systems.

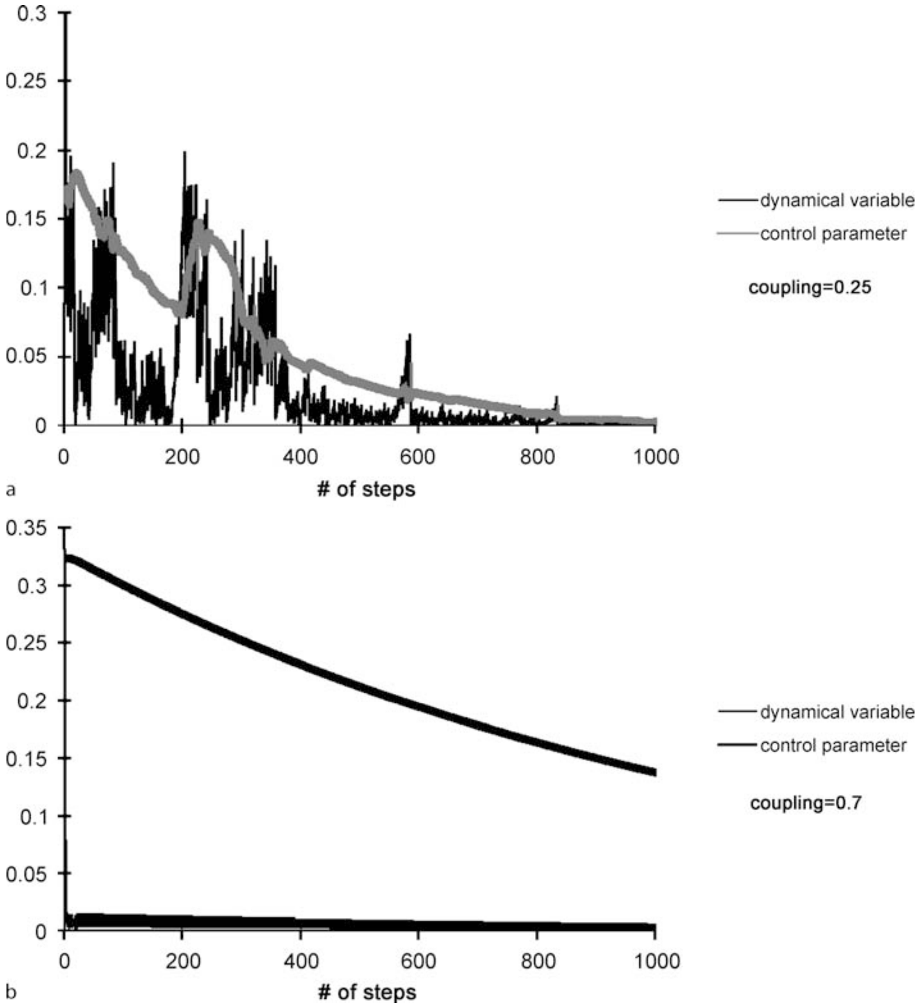
The results demonstrated that behavioral synchronization increased both with increases in α and in the similarity in r. At each degree of coupling, synchronization increased with greater similarity in partner's internal states. Likewise, at each level of similarity in internal states, synchronization increased with increased strength of coupling. These results have straightforward implications. If two people have similar control parameters (internal states), relatively little coupling (e. g., influence, communication, mutual reinforcement, monitoring) is necessary for them to achieve a high degree of synchronization in their behavior. If the partners have different internal states, on the other hand, high mutual influence is required to maintain the same level of synchronization. In a close relationship, this suggests that constant and intense communication may be a sign that the partners are not well coordinated with respect to relevant internal states (e. g., temperament, desires values). But when the partners are similar with respect to such internal states, they can devote their energy to common pursuits rather than to constant clarification, monitoring, and other forms of influence.

Modeling Internal Synchronization

Some internal parameters, such as moods and emotional states, vary considerably across time and settings. If a particular context induces a common mood or emotion (e. g., joy or sadness) in a dyad or group, interpersonal synchronization is easy to achieve. Other internal states, however, demonstrate greater stability and less likely to vary across different contexts. Attitudes, values, and personality traits, in particular, are commonly considered to be enduring (cross-situational) features of a person's psychological make-up. Yet, even these internal parameters admit to variability and even modification. The Nowak et al. [82,83] model has been used to understand and investigate the nature of such change. The core assumption is that individuals are motivated to achieve coordination in their internal parameters [13,62]. To satisfy this motive, individuals vary their internal parameters in a direction that leads to increasing synchronization. When synchronization is achieved, the value of the resultant control parameter is engraved as an attractor for that internal state. In essence, people are assumed to develop stable internal states through social synchronization (e. g. [112]).

To model this process, Nowak et al. simply assumed that the value of each individual's control parameter drifts somewhat in the direction of the value of the other individual's control parameter on each simulation step. How quickly the respective control parameters begin to match depends on the size of the initial discrepancy between them and on the rate at which they drift. This process occurs with each individual knowing the exact value of the other individual's control parameter. This is an important feature of the model, since experimental research has shown that people's internal states are often difficult to infer [43,47,71,113]. The model assumes only that each individual remembers the other individual's most recent behaviors (i. e., the most recent values of x) as well as his or her own most recent behaviors. Each individual compares his or her own behavior with that of the other individual, and then adjusts his or her internal parameter in order to promote increased similarity with the other person's behavior pattern, until a match is achieved [115]. Thus, if the other individual's behavior is more complex than the individual's own pattern of behavior, the individual slightly increases the value of his or her own control parameter. Conversely, the individual slightly decreases the value of his or her own control parameter if the other individual's behavior is less complex than his or her own. In short, interacting individuals estimate one another's internal state by monitoring the evolution of one another's behavior.

Nowak et al. [82,83] ran simulations to investigate the convergence of behavior and internal states under both relatively weak and relatively strong coupling ($\alpha = .25$ and .7, respectively). The y-axis in Fig. 3 corresponds to the magnitude of difference between the two systems in their behavior or internal states, and the x-axis corre-



Social Psychology, Applications of Complexity to, Figure 3

a Convergence of behavior and internal states under weak coupling b Convergence of behavior and internal states under strong coupling [82]

sponds to time, as reflected in the number of iterations. Consider first the results observed when the coupling was weak (Fig. 3a). The behavior of the two systems converged in a relatively slow and non-linear manner. The control parameters also showed a clear tendency towards convergence. When a match in internal states was achieved, moreover, full synchronization of behavior was also obtained. Compare these results with those observed under strong coupling (Fig. 3b). Note that although there was immediate synchronization of behavior, the control parameters failed to synchronize, even after 1000 iterations.

The differential results obtained for weak and strong coupling may have noteworthy implications for interpersonal relations. First, even for people with very different internal parameters, strong coupling tends to promote full synchronization of behavior. Once synchronization is achieved, the two people may be totally unaware that their internal states are different. This suggests that if the coupling were to be reduced in magnitude (or removed altogether), the dynamics of the two individuals would immediately diverge. Hence, people who employ very strong influence (e. g., reinforcement, monitoring) to obtain coordination of behavior may effectively hinder synchronization of their respective internal parameters. In more general terms, there may be an optimal level of influence and control over one another's behavior in interpersonal relations [111]. When influence is too weak, synchronization between individuals may fail to develop. But when influence is very strong, it can prevent the development of a relationship based on mutual understanding and empathy. Intermediate levels of mutual influence, then, may be most effective for the development of synchronization on a deep level. Stated differently, the most advantageous degree of coupling (e. g., influence) is the minimal amount necessary to achieve synchronization.

These implications are consistent with extensive research in social psychology suggesting that behavior attributed to external causes is less likely to promote psychological change than is behavior attributed to internal causes. For example, people are resistant to changing their preferences and attitudes if they believe that their behavior is in response to direct orders, rewards, threats, and other external influences [7,56]. Salient external influences, in fact, may activate mechanisms to counter the influences, creating an internal state that is opposite of the intended effect of the influence [8,111].

Future Directions

The application of complexity science to social psychology makes for an ironic discipline. The tools of complexity and dynamical systems are ideally suited to capture the subtlety and uniqueness of human experience, yet they are grounded in concepts and methods that provide meaningful integration with the natural sciences. Disciplines such as mathematics, physics, and chemistry have affirmed the significance of intrinsic dynamics, nonlinear phenomena, self-organization, and complexity. In recent years, psychologists have become increasingly cognizant that these features of systems in nature have counterparts in mental, interpersonal, and collective experience. It would be odd for a contemporary social psychologist to discount the potential for emergence or to ignore the role of computer simulations and time series in illuminating how minds, groups, and societies work.

A word of caution is in order, however. Social reality is not the same as physical reality. Unlike atoms, individuals are not interchangeable, and groups are more than self-organized ensembles of simple particles. People have values and beliefs, universal concerns and idiosyncratic tendencies, and moments of self-reflection and sudden impulse. One of the basic rules of human psychology, moreover, is the capacity for reflecting on one's operating rules and attempting to override them. People exist in a symbolically constructed world and do not respond in a reflexive way to objective reality. These unique features of human experience cannot be chalked up to the recognition that people are dynamic and complex. The task of dynamical social psychology is thus more daunting than discovering the differential equations that govern interpersonal and collective dynamics. Models derived from complexity and dynamical systems provide a foundational science for the discipline and much remains to be discovered within this new paradigm. Ultimately, though, the properties that separate us from other systems in nature must be incorporated into theoretical models. It is an ironic testament to the complexity perspective that a coherent theory of social psychology should be assembled from elements that are both universal within nature and unique to human experience.

Bibliography

Primary Literature

1. Abelson RP (1979) Social clusters and opinion clusters. In: Holland PW, Leinhardt S (eds) Perspectives in social network research. Academic Press, New York, pp 239–256
2. Abraham RH, Shaw CD (1992) Dynamics, the geometry of behavior, 2nd edn. Addison-Wesley, Reading
3. Allport GW (1968) The historical background of modern social psychology. In: Lindzey GA, Aronson E (eds) The Handbook of Social Psychology, vol 1. Addison-Wesley, Reading

4. Asch SE (1946) Forming impressions of personalities. J Abnorm Soc Psychol 41:258–290
5. Baron RM, Amazeen PM, Beek PJ (1994) Local and global dynamics of social relations. In: Vallacher RR, Nowak A (eds) Dynamical systems in social psychology. Academic Press, San Diego, pp 111–138
6. Beek PJ, Hopkins B (1992) Four requirements for a dynamical systems approach to the development of social coordination. Huma Mov Sci 11:425–442
7. Bem DJ (1967) Self-perception: An alternative interpretation of cognitive dissonance phenomena. Psychol Rev 74:183–200
8. Brehm SS, Brehm JW (1981) Psychological reactance: A theory of freedom and control. Academic Press, New York
9. Brown KW, Moskowitz DS (1998) Dynamic stability of behavior: The rhythms of our interpersonal lives. J Personal 66:105–134
10. Bugental DB (2000) Acquisition of the algorithms of social life: A domain-base approach. Psychol Bull 126:187–219
11. Buss DM (1999) Evolutionary psychology: The new science of the mind. Allyn Bacon, Boston
12. Cannon WB (1932) The wisdom of the body. Norton, New York
13. Caporeal LR, Baron RM (1997) Groups as the mind's natural environment. In: Simpson JA (ed) Evolutionary social psychology. Lawrence Erlbaum Associates, Hillsdale, pp 317–344
14. Carver CS, Scheier MF (1999) Themes and issues in the self-regulation of behavior. In: Wyer RS Jr (ed) Advances in social cognition, vol 12. Erlbaum, Mahwah, pp 1–105
15. Carver CS, Scheier MF (2002) Control processes and self-organization as complementary principles underlying behavior. Personal Soc Psychol Rev 6:304–315
16. Chartrand TL, Bargh JA (1999) The chameleon effect: The perception-action link and social interaction. J Pers Soc Psychol 76:893–910
17. Coleman PT (2003) Characteristics of protracted, intractable conflict: Towards the development of a meta-framework - I. First paper in a three-paper series. Peace Confl: J Peace Psychol 9:1–37
18. Coleman PT, Bui-Wrzosinska L, Vallacher RR, Nowak A (2006) Protracted conflicts as dynamical systems. In: Schneider AK, Honeyman C (eds) The negotiator's fieldbook: The desk reference for the experience negotiator. American Bar Association Books, Chicago, pp 61–74
19. Coleman PT, Vallacher RR, Nowak A, Bui-Wrzosinska L (2007) Intractable conflict as an attractor: A dynamical systems approach to conflict escalation and intractability. Am Behav Sci 50:1454–1475
20. Condon WS, Ogston WD (1967) A segmentation of behavior. J Psychiatr Res 5:221–235
21. Cooley CH (1902) Human nature and the social order. Scribner, New York
22. Deutsch M (1973) The resolution of conflict: Constructive and destructive processes. Yale University Press, New Haven
23. Dittman AT, Llewellyn LG (1969) Body movement and speech rhythm in social conversation. J Personal Soc Psychol 11:98–106
24. Durkheim E (1938) The rules of sociological method. University of Chicago Press, Chicago
25. Eckmann JP, Ruelle D (1985) Ergodic theory of chaos and strange attractors. Rev Mod Phys 57:617–656
26. Eiser JR (1994) Attitudes, chaos, and the connectionist mind. Blackwell, Oxford
27. Feigenbaum MJ (1978) Quantitative universality for a class of nonlinear transformations. J Stat Phys 19:25–52
28. Festinger L, Schachter S, Back K (1950) Social pressures in informal groups. Stanford University Press, Stanford
29. Fiske AP (1992) The four elementary forms of sociality: Framework for a unified theory of social relations. Psychol Rev 99:689–723
30. Glass L, Mackey MC (1988) From clocks to chaos: The rhythms of life. Princeton University Press, Princeton
31. Goldstein J (1996) Causality and emergence in chaos and complexity theories. In: Sulis W, Combs A (eds) Nonlinear dynamics and human behavior. World Scientific, Singapore, pp 161–190
32. Gottman JM (1979) Detecting cyclicity in social interaction. Psychol Bull 86:338–348
33. Gottman J, Swanson C, Swanson K (2002) A general systems theory of marriage: Nonlinear difference equation modeling of marital interaction. Personal Soc Psychol Rev 4:326–340
34. Gottman JM, Murray JD, Swanson CC, Tyson R, Swanson KR (2002) The Mathematics of Marriage. MIT Press, Cambridge
35. Guastello SJ (1995) Chaos, catastrophe, and human affairs: Applications of nonlinear dynamics to work, organizations, and social evolution. Erlbaum, Mahwah
36. Guastello SJ, Pincus D, Gunderson PR (2006) Electrodermal arousal between participants in a conversation: Nonlinear dynamics and linkage effects. Nonlinear Dyn Psychol Life Sci 10:365–399
37. Higgins ET (1987) Self-discrepancy: A theory relating self and affect. Psychol Rev 94:319–340
38. Hock HS, Kelso JAS, Schoner G (1993) Bistability and hysteresis in the organization of apparent motion pattern. J Exp Psychol Hum Percept Perform 19:63–80
39. Holland JH (1995) Emergence: From chaos to order. Addison-Wesley, Reading
40. Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences USA 79:2554-2558
41. James W (1890) Principles of psychology. Holt, New York
42. Johnson SL, Nowak A (2002) Dynamical patterns in bipolar depression. Personal Soc Psychol Rev 6:380–387
43. Jones EE, Davis KE (1965) From acts to dispositions: The attribution process in person perception. In: Berkowitz L (ed) Advances in experimental social psychology, vol 2. Academic Press, New York, pp 220–266
44. Kaneko K (ed) (1993) Theory and applications of coupled map lattices. World Scientific, Singapore
45. Kaplowitz SA, Fink EL (1992) Dynamics of attitude change. In: Levine RL, Fitzgerald HE (eds) Analysis of dynamic psychological systems, vol 2. Plenum, New York, pp 341–369
46. Kelley HH (1967) Attribution in social psychology. Neb Symp Motiv 15:419–422
47. Kelley HH (1971) Attribution in social interaction. Morristown
48. Kelso JAS (1995) Dynamic patterns: The self-organization of brain and behavior. The MIT Press, Cambridge
49. Kenrick DT, Maner JK, Butner J, Li NP, Becker DV, Schaller M (2002) Dynamical evolutionary psychology: Mapping the domains of the new interactionist paradigm. Personal Soc Psychol Rev 6:347–356

50. Larsen RJ (1987) The stability of mood variability: A spectral analytic approach to daily mood assessments. J Personal Soc Psychol 52:1195–1204
51. Larsen RJ, Kasimatis M (1990) Individual differences in entrainment of mood to the weekly calendar. J Personal Soc Psychol 58:164–171
52. Latané B (1981) The psychology of social impact. Am Psychol 36:343–356
53. Latané B, Nowak A (1994) Attitudes as catastrophes: From dimensions to categories with increasing involvement. In: Vallacher RR, Nowak A (eds) Dynamical systems in social psychology. Academic Press, San Diego, pp 219–249
54. Latané B, Nowak A (1997) The causes of polarization and clustering in social groups. Prog Commun Sci 13:43–75
55. Latané B, Nowak A, Liu J (1994) Measuring emergent social phenomena: dynamism, polarization and clustering as order parameters of social systems. Behav Sci 39:1–24
56. Lepper MR, Greene D (eds) (1978) The hidden costs of reward. Erlbaum, Hillsdale
57. Lewis MD (2005) Bridging emotion theory and neurobiology through dynamic systems modeling. Behav Brain Sci 28:169–194
58. Lewenstein M, Nowak A, Latané B (1993) Statistical mechanics of social impact. Phys Rev 45:703–716
59. Lewin K (1936) Principles of topological psychology. McGraw-Hill, New York
60. Linville PW (1985) Self-complexity and affective extremity: Don't put all your eggs in one cognitive basket. Soc Cogn 3:94–120
61. Markus H, Nurius P (1986) Possible selves. Am Psychol 41:954–969
62. Marsh KL, Richardson MJ, Baron RM (2006) Contrasting approaches to perceiving and acting with others. Ecol Psychol 18:1–37
63. Mazanov J, Byrne DG (2006) A cusp catastrophe model analysis of changes in adolescent substance use: Assessment of behavioural intention as a bifurcation variable. Nonlinear Dyn Psychol Life Sci 10:445–470
64. McGrath JE, Kelley JR (1986) Time and human interaction: Toward a psychology of time. Guilford Publications, New York
65. Mead GH (1934) Mind, self, and society. University of Chicago Press, Chicago
66. Miller NE (1944) Experimental studies of conflict. In: Hunt JM (ed) Personality and the behavior disorders. Ronald, New York
67. Moscovici S, Lage E, Naffrechoux M (1969) Influence of a consistent minority on responses of a majority in a color perception task. Sociometry 32:365–379
68. Myers DG, Lamm H (1976) The group polarization phenomenon. Psychol Bull 83:602–627
69. Newtson D (1994) The perception and coupling of behavior waves. In: Vallacher RR, Nowak A (eds) Dynamical systems in social psychology. Academic Press, San Diego, pp 139–167
70. Nezlek JB (1993) The stability of social interaction. J Personal Soc Psychol 65:930–941
71. Nisbett R, Ross L (1980) Human inference: Stategies and shortcomings of social judgment. Prentice-Hall, Englewood Cliffs
72. Nowak A (2004) Dynamical minimalism: Why less is more in psychology. Personality and Social Psychology Review 8:183-192
73. Nowak A, Lewenstein M (1994) Dynamical systems: A tool for social psychology? In: Vallacher RR, Nowak A (eds) Dynamical systems in social psychology. Academic Press, San Diego, pp 17–53
74. Nowak A, Vallacher RR (1998) Dynamical social psychology. Guilford, New York
75. Nowak A, Vallacher RR (2001) Societal transition: Toward a dynamical model of social change. In: Wosinska W, Cialdini RB, Barrett DW, Reykowski J (eds) The practice of social influence in multiple cultures. Lawrence Erlbaum, Mahwah, pp 151–171
76. Nowak A, Szamrej J, Latané B (1990) From private attitude to public opinion: A dynamic theory of social impact. Psychol Rev 97:362–376
77. Nowak A, Lewenstein M, Szamrej J (1993) Social transitions occur through bubbles. Sci Am Pol vers 12:16–25
78. Nowak A, Urbaniak J, Zienkowski L (1994) Clustering processes in economic transition. RECESS Res Bull 3:43–61
79. Nowak A, Lewenstein M, Frejlak P (1996) Dynamics of public opinion and social change. In: Hegselman R, Pietgen HO (eds) Modeling social dynamics: Order, chaos, and complexity. Helbin, Vienna, pp 54–78
80. Nowak A, Vallacher RR, Borkowski W (2000) Modeling the temporal coordination of behavior and internal states. In: Ballot G, Weisbuch G (eds) Applications of simulation to the social sciences. Hermes Science Publications, Oxford, pp 67–86
81. Nowak A, Vallacher RR, Tesser A, Borkowski W (2000) Society of self: The emergence of collective properties in self-structure. Psychol Rev 107:39–61
82. Nowak A, Vallacher RR, Zochowski M (2002) The emergence of personality: Personal stability through interpersonal synchronization. In: Cervone D, Mischel W (eds) Advances in personality science. Guilford Press, New York, pp 292–331
83. Nowak A, Vallacher RR, Zochowski M (2005) The emergence of personality: Dynamic foundations of individual variation. Dev Rev 25:351–385
84. Nowak A, Vallacher RR, Kus M, Urbaniak J (2005) The dynamics of societal transition: Modeling nonlinear change in the Polish economic system. Int J Sociol 35:65–88
85. Nowak A, Vallacher RR, Bui-Wrzosinska L, Coleman PT (2007) Attracted to conflict: A dynamical perspective on malignant social relations. In: Golec A, Skarzynska K (eds) Understanding social change: Political psychology in Poland. Nova Science Publishers, Haauppague
86. Read SJ, Miller LC (2002) Virtual personalities: A neural network model of personality. Personal Soc Psychol Rev 6:357–369
87. Rosenblum LD, Turvey MT (1988) Maintenance tendency in coordinated rhythmic movements: Relative fluctuations and phase. Neuroscience 27:289–300
88. Schachter S (1951) Deviation, rejection and communication. J Abnorm Soc Psychol 46:199–207
89. Schuster HG (1984) Deterministic chaos. Physik, Vienna
90. Sherif M (1936) The psychology of social norms. Harper, New York
91. Shinbrot T (1994) Synchronization of coupled maps and stable windows. Phys Rev E 50:3230–3233
92. Shoda Y, LeeTiernan S, Mischel W (2002) Personality as a dynamical system: Emergence of stability and distinctiveness

from intra- and interpersonal interactions. Personal Soc Psychol Revi 6:316–325

93. Showers CJ (1995) The evaluative organization of self-knowledge. In: Kernis MH (ed) Efficacy, agency, and self-esteem. Plenum, New York, pp 101-120
94. Simon D, Holyoak KJ (2002) Structural dynamics of cognition: From consistency theories to constraint satisfaction. Personal Soc Psychol Rev 6:283–294
95. Swann WB Jr (1990) To be adored or to be known? The interplay of self-enhancement and self-verification. In: Higgins ET, Sorrentino RM (eds) Handbook of motivation and cognition: Foundations of social behavior, vol 2. Guilford, New York, pp 408–448
96. Swann WB, Hixon JG, Stein-Seroussi A, Gilbert D (1990) The fleeting gleam of praise: Cognitive processes underlying behavioral reactions to self-relevant feedback. J Personal Soc Psychol 59:17–26
97. Tesser A (1978) Self-generated attitude change. In: Berkowitz L (ed) Advances in experimental social psychology, vol 11. Academic Press, New York, pp 85–117
98. Tesser A, Martin L, Cornell D (1996) On the substitutability of self-protective mechanisms. In: Gollwitze PM, Bargh JA (eds) The psychology of action. Guilford Publications, New York, pp 48–68
99. Thagard P, Nerb J (2002) Emotional Gestalts: Appraisal, change, and the dynamics of affect. Personal Soc Psychol Rev 6:274–282
100. Tickle-Degnen L, Rosenthal R (1987) Group rapport and nonverbal behavior. Rev Personal Soc Psychol 9:113–136
101. Turvey MT (1990) Coordination. Am Psychol 4:938–953
102. Vallacher RR, Nowak A (eds) (1994) Dynamical systems in social psychology. Academic Press, San Diego
103. Vallacher RR, Nowak A (1994) The stream of social judgment. In: Vallacher RR, Nowak A (eds) Dynamical systems in social psychology. Academic Press, San Diego, pp 251–277
104. Vallacher RR, Nowak A (1997) The emergence of dynamical social psychology. Psychol Inq 8:73–99
105. Vallacher RR, Nowak A (1999) The dynamics of self-regulation. In: Wyer RS Jr (ed) Advances in self-regulation, vol 12. Lawrence Erlbaum, Mahwah, pp 241–259
106. Vallacher RR, Nowak A (2007) Dynamical social psychology: Finding order in the flow of human experience. In: Kruglanski AW, Higgins ET (eds) Social psychology: Handbook of basic principles, 2nd edn. Guilford Publications, New York, pp 734–758
107. Vallacher RR, Wegner DM (1987) What do people think they're doing? Action identification and human behavior. Psychol Rev 94:1–15
108. Vallacher RR, Wegner DM (1989) Levels of personal agency: Individual variation in action identification. J Personal Soc Psychol 57:660–671
109. Vallacher RR, Nowak A, Kaufman J (1994) Intrinsic dynamics of social judgment. J Personal Soc Psychol 66:20–34
110. Vallacher RR, Nowak A, Froehlich M, Rockloff M (2002) The dynamics of self-evaluation. Personal Soc Psychol Rev 6:370–379
111. Vallacher RR, Nowak A, Miller ME (2003) Social influence and group dynamics. In: Weiner I, Millon T, Lerner MJ (eds) Handbook of psychology, vol 5. Personality and social psychology. Wiley, New York, pp 383–417
112. Vallacher RR, Nowak A, Zochowski M (2007) Dynamics of social coordination: The synchronization of internal states in close relationships. In: Hauf P, Forsterling F (eds) Making minds: The shaping of human minds through social context. John Benjamins Publishing Company, Amsterdam, pp 31–46
113. Wegner DM, Vallacher RR (1977) Implicit psychology: An introduction to social cognition. Oxford University Press, New York
114. Wegner DM, Vallacher RR, Kiersted G, Dizadji D (1986) Action identification in the emergence of social behavior. Soc Cogn 4:18–38
115. Zochowski M, Liebovitch LS (1999) Self-organizing dynamics of coupled map systems. Phys Rev E 59:2830

Books and Reviews

Guastello S, Koopmans M, Pincus D (eds) (2008) Chaos and Complexity in Psychology: The Theory of Nonlinear Dynamical Systems. Cambridge University Press, New York

Hegselman R, Pietgen HO (eds) (1996) Modeling social dynamics: Order, chaos, and complexity. Helbin, Vienna

Lewin K (1936) Principles of topological psychology. McGraw-Hill, New York

Liebrand W, Nowak A, Hegselman R (eds) (1998) Computer modeling and the study of dynamic social processes. Sage, New York

Read SJ, Miller LC (eds) (1998) Connectionist models of social reasoning and social behavior. Erlbaum, Mahwah

Smith ER (1996) What do connectionism and social psychology offer each other? J Personal Soc Psychol 70:893–912

Vallacher RR, Read SJ, Nowak A (eds) (2002) The dynamical perspective in social psychology. Personal Soc Psychol Rev, Special Issue 6:264–387

Soft Computing, Introduction to

Janusz Kacprzyk
Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

As science progresses, scholars and researchers continue to develop models of various phenomena and problems. These models have become more and more formal, notably mathematical, and hence have relied more and more on computing.

The initial period has been characterized by computing meant as well-defined algorithmic procedures implemented by using computing architectures, tools and techniques based on traditional binary (yes – no or 0 – 1) logic. This logic has also been a basis for all kinds of approaches within the field of artificial intelligence, a field of science and technology that has been vividly developed since the mid-1900s and aimed at devising machines capable of performing human-specific tasks that require intelligence. Symbolic computation has become a main paradigm within artificial intelligence, and the role of numerical computing has been neglected to a large extent.

As promises of artificial research gurus that machines will be able to perform virtually all sophisticated human-specific tasks and practically replace humans in a short time have failed, it has become more and more evident that sticking to symbolic computations only is a mistake. Clearly, there may be tasks for which symbolic computations are adequate, but there may be tasks, presumably even more, when numerical computations are indispensable.

Even if this necessity of a synergistic use of symbolic and numerical computations in the construction of intelligent systems has become obvious, people have been, for a long time, not fully convinced that, since the human being is a crucial element in all kinds of research, analysis and construction of broadly perceived intelligent systems, computational techniques and processes involved therein should reflect human-specific features.

In our context, the main feature of the human element in computation is that natural language is the only fully natural means of articulation and communication of a human being, and natural language is plagued by inherent vagueness, ambiguity, imprecision, etc. These kinds of information imperfections are out of scope of conventional mathematical tools based on – for instance – probabilistic or statistical tools.

The necessity of developing a new framework which would differ from the traditional "hard" computing based on traditional mathematics, i. e., basically on binary logic, and the dichotomy of true and false, has been recognized quite early. However, it was Lotfi A. Zadeh from the University of California at Berkeley who first pronounced that necessity – in the context of the analysis of complex animate systems, notably human-centered – in a more comprehensive way in the early 1960s. In 1965 he introduced fuzzy sets theory, presumably the first comprehensible, yet simple and efficient calculus of imprecision. A fuzzy set is a class to which elements can belong to a degree, from full membership, through all intermediate values, to full non-membership. Clearly, this has provided a constructive means for the representation of imprecisely specified concepts, relations, reasoning schemes, etc. Zadeh then proceeded to the formulation of possibility theory, and finally based his ideas on a more direct link to natural language which has culminated in his computing with words and perceptions.

The introduction of a fuzzy set, followed by numerous applications of fuzzy logic-based tools and techniques, notably for control, ranging from the control of domestic appliances, automatic transmissions in cars, to the control of industrial installations, has certainly been very important for the proliferation of fuzzy logic, maybe even more generally for non-conventional computational tools and techniques.

As a next step, Zadeh has advocated a wider view of those new computational tools and techniques and coined the term of *soft computing*. Basically, soft computing may be viewed as a departure from conventional (hard and "precise") computing in that it is tolerant of imprecision, uncertainty, partial truth, and approximate relationships, approximately satisfied properties, etc.

The main principle underlying soft computing is to be able to exploit a tolerance for imprecision, uncertainty, partial truth, and approximation to attain tractable, robust and computationally inexpensive modeling and solution tools and techniques. A natural connection to neural computing and evolutionary computing has more recently been recognized.

It is usually assumed that the principal components of soft computing are fuzzy logic (complemented with rough sets theory which may be viewed to address related problems of imprecise information but understood and dealt with in a different way), neural computation, and evolutionary computation. These areas are sometimes complemented with machine learning and probabilistic reasoning, with the latter subsuming belief networks, chaos theory and parts of learning theory. Moreover, for a long time, rough sets have been considered to be a relevant part of soft computing.

However, it should be noticed that soft computing is not a confluence of more or less independent fields, but is meant as a synergistic combination of various methodologies and methods in which each one contributes in a synergistic and complementary way to the analysis and solution of a problem.

In this section we will present the main components of soft computing. The concepts of synergy and complementarity mentioned above will provide a pivotal perspective.

We start with a brief introduction to fuzzy sets and systems (see J. Kacprzyk, ► Fuzzy Sets Theory, Foundations of). First, we define the concept of a fuzzy set followed by the definitions of its main properties. The idea of a fuzzy relation is then introduced, and shown to be a convenient apparatus for devising approximate reasoning schemes. The concept of a linguistic variable is shown to be a basis for a natural-language-based approach to modeling and reasoning. Fuzzy arithmetic is presented making it possible to use imprecisely specified (fuzzy) numbers in models and algorithms. A universal problem of decision making is dealt with under fuzzy goals, constraints, etc., in a static and dynamic context.

Though fuzzy sets in the traditional Zadeh sense do provide a simple yet efficient means for the representation and handling of imprecision, some natural extensions of the basic concept of a fuzzy set have appeared. One of the most relevant and popular is the concept of a type 2 fuzzy set in which the degree of membership itself, which is a real number from the unit interval in the traditional fuzzy set, is now a fuzzy set defined in the unit interval. This natural extension has implied relevant extensions of fuzzy modeling and reasoning tools and techniques which are presented in detail, and potentials for applications are underlined (see R.I. John and J.M. Mendel, ▶ Fuzzy Logic, Type-2 and Uncertainty).

As in any formal model in which, as is true in virtually all cases in reality, multiple aspects, points of view and entities have to be accounted for, a proper aggregation is of a paramount importance. A comprehensive account of various aggregation operators is provided, both from a strictly mathematical point of view and from a more constructive, application oriented one (see V. Torra, ▶ Aggregation Operators and Soft Computing).

The discussion of the foundations of fuzzy logic is completed with a brief account of possibility theory initiated by Zadeh in the late 1970s as an extension of fuzzy sets theory and fuzzy logic, a mathematical theory for dealing with certain types of imperfect (uncertain) information in a way that is an alternative to probability theory. Main concepts related to possibility theory are presented, and a historical perspective is provided (see D. Dubois and H. Prade, ▶ Possibility Theory).

The comprehensive presentation of foundational concepts and properties related to fuzzy logic is then followed by sections presenting some more relevant areas which can be of use in the analysis of many systems and in the solution of many problems in diverse areas of science and technology.

First, fuzzy optimization and fuzzy mathematical programming are outlined starting with basic concepts of optimality, feasibility, etc., under fuzzy information. Then, various classes of fuzzy optimization and fuzzy mathematical programming problem classes are briefly outlined and basic solution concepts and algorithms are presented. Some examples of applications are mentioned (see W. Lodwick and E. Untiedt, ▶ Fuzzy Optimization).

The topic of statistics with imprecise data is considered next, a very interesting topic from a conceptual point of view, and practically relevant. First, a sound motivation is presented by pointing out that, in reality, much available data is imprecise, and so they cannot be used by powerful and well founded, yet too "hard" traditional statistical tools and techniques. New concepts are formulated and extensions of well known statistical means under an imprecise information are presented. Some applications are outlined (see M.A. Gil and O. Hryniewicz, ▶ Statistics with Imprecise Data).

The next papers are concerned with the two other key elements of soft computing: neural computation and evolutionary computation. To emphasize their synergy and complementarity, these tools have been presented in a hybrid context, that is as a presentation of new fields in which the confluence of fuzzy, neural and evolutionary tools yields a new quality, and results in the emergence of a new class of tools and techniques which have already changed the landscape of soft computing.

We start with the neuro-fuzzy systems, which refer to a combination of (artificial) neural networks and fuzzy logic, and may be viewed as a main, fundamental step into the development of hybrid intelligent systems (see L. Rutkowski, K. Cpałka, R. Nowicki, A. Pokropińska, R. Scherer, ▶ Neuro-fuzzy Systems). Basically, neuro-fuzzy systems try to combine transparent, human-like reasoning types of fuzzy rule-based systems with a parallel, connectionist type of neural networks. Therefore, they are universal approximators with an ability to yield interpretable IF-THEN rules, bridging the gap between accuracy and interpretability. A broad coverage of main architectures and implementations is presented.

Though neuro-fuzzy systems have been a considerable step forward towards in developing modeling tools of a better expressive power, effectiveness and efficiency, some natural extensions have been proposed that take advantage of the increasing popularity of evolutionary computation. It has been shown theoretically and experimentally that the incorporation in broadly perceived fuzzy systems, including neuron-fuzzy systems, of a learning mechanism which can make it possible to adjust the systems' structure and parameters may greatly enhance performance. The use of evolutionary computation has led in this context to the concept of an evolving fuzzy system that has been considered a promising alternative for some time (see P. Angelov, ▶ Evolving Fuzzy Systems).

The above mentioned hybridization techniques, which basically boil down to architectures and algorithms being a synergistic combination of fuzzy systems, notably fuzzy rule based systems, neural networks and evolutionary computation, in particular evolutionary learning, can be further extended when applied to real world problems which always provide much inspiration. Applications for systems modeling and control are of a universal importance for many fields of science and technology and hence are dealt with in detail (see O. Castillo and P. Melin, ► Hybrid Soft Computing Models for Systems Modeling and Control).

The chapters briefly described above have been related to fuzzy sets and some of their hybridizations by including elements of neural networks and evolutionary computation to attain some "superadditivity" of the strengths and power of the particular single tools and techniques.

Rough sets theory, introduced by the late Zdzislaw Pawlak in the early 1980s, has been for some time considered a very promising tool for dealing with imperfect information, notably in data analysis and decision making.

Basically, a rough set is a formal approximation of a conventional crisp set in that it is represented as a pair of two traditional crisp sets which constitute the lower and the upper approximation of the original. These can be viewed as sets of elements that possibly and surely belong to the original set. Rough sets theory is presented in much detail, first in a more traditional, pure setting in Pawlak's spirit. Various concepts are presented, relations between them are formulated and proved, and their possible applications are outlined, notably in the relevant areas of data mining and knowledge discovery. Some extensions of the basic concept of a rough set are also presented (see J. Peters, A. Skowron and J. Stepaniuk, ► Rough Sets: Foundations and Perspectives).

Next, a "meta-problem" in science, that is decision making, is dealt with in the context of rough sets. In addition to the basic definition of a rough set, which provides much insight into the formalization and solution of various classes of decision making problems, a novel idea of a dominance-based rough set is presented. Its use to derive new, more human consistent solution concepts in decision making and data analysis is proposed (see R. Słowiński, S. Greco and B. Matarazzo, ► Rough Sets in Decision Making).

To summarize, we have tried to present the vast and diversified area of soft computing in a comprehensive, illustrative and constructive way. The basic perspective assumed may be viewed as presenting soft computing as a significant paradigm shift in computing in that it tries to exploit the fact that the human mind, unlike the computer we have now, has a remarkable ability to store and process information which is predominantly imprecise, uncertain and granular.

We have started with some basic fundamental concepts and properties, and then have devoted much space to the presentation of what is characteristic for soft computing: a synergistic, complementary use of various tools and techniques to solve problems in an effective and efficient way. This has led to various types of hybridization in which tools from fuzzy systems, rough sets, neural networks and evolutionary computation have been employed. It seems that soft computing should play a more and more important role in the years to come, in which more and more emphasis should be put on human-consistent and human-centric tools and techniques, and on hybrid systems.

Software Architectures for Autonomy

ROBERT A. TOUCHTON[1], CARL D. CRANE III[2]
[1] Honeywell International, Phoenix, USA
[2] University of Florida, Gainesville, USA

Article Outline

Glossary

Reference architecture (RA) A reference architecture is a framework containing valuable implementation guidance for meeting the requirements of an enterprise.

Automated guided vehicle (AGV) A vehicle that can be programmed to automatically drive to designated points and perform preprogrammed functions.

JAUS The Joint Architecture for Unmanned Systems (JAUS) Reference Architecture defines a set of reusable components and their interfaces. The architecture emphasizes vehicle platform independence, mission isolation, computer hardware independence, and technology independence.

NIST 4D/RCS The acronym NIST 4D/RCS refers to the National Institute of Standards and Technology Real-time Control System. 4D/RCS describes in detail the functions and associated interfaces necessary to provide sensory processing, world modeling, knowledge management, cost/benefit analysis, and behavior generation.

Adaptive planning framework The Adaptive Planning Framework addresses dynamic situation assessment, behavior management, and decision-making. It incorporates a three-stage process of 1) understanding the current situation, 2) understanding the suitability and viability of the available behaviors in light of that situation, and 3) providing the capability to autonomously make and execute behavior-related decisions, all in real-time.

Service oriented architecture (SOA) A Service Oriented Architecture maintains a strictly enforced standardized interface among entities. A standardized messaging construct enables one entity to request a service from another entity and for that service provider to send its response.

Definition of the Subject

Developing an autonomous system, such as an autonomous ground vehicle, that can operate and maneuver in an unstructured environment is a complicated task. One of the most daunting issues facing autonomous vehicle researchers is how to best exploit sensor and other information discovered during the execution of a plan. If the system takes too long to deliberate on the possible meanings and implications of this newfound data and knowledge, the vehicle may well have progressed beyond the point where it can benefit from it. Indeed, it may now be sitting atop the unforeseen obstacle that spawned the influx of new information that was being processed.

The execution of specific autonomous behaviors is becoming reasonably well understood, such as “waypoint-following with obstacle detection”, though improvements and breakthroughs in these areas continue. However, the autonomous selection of which behavior(s) should be invoked, and in what sequence and by what method, is in need of movement in the state of the practice. Advanced ways of thinking about, organizing, and applying situational knowledge to macro-level planning and decision-making are needed by the autonomous robotics community in order to achieve the full potential of the field. This article discusses various approaches and various reference architectures that have been developed to address these problems.

Introduction

Three predominant standards are discussed in the following sections, i. e. the Joint Architecture for Unmanned Systems (JAUS) Reference Architecture, the National Institute of Standards and Technology (NIST) 4D/RCS (Real-time Control System), and the Service Oriented Architecture. A fourth architecture referred to as the Distributed Architecture for Mobile Navigation (DAMN) is also discussed although this architecture is not widely adopted. Lastly, the Adaptive Planning Framework approach that was developed at the University of Florida to address the problem of decision making is introduced.

Joint Architecture for Unmanned Systems (JAUS)

The Joint Architecture for Unmanned Systems (JAUS) Reference Architecture, Version 3.2 [18] defines a set of reusable components and their interfaces. In order to ensure that the architecture will be applicable to the entire domain of unmanned mobile systems, the following four characteristics have been considered by the JAUS Working Group in the creation of the Reference Architecture:

1. Vehicle platform independence. In order for JAUS components to be interoperable, no assumptions about the underlying vehicle or its means of propulsion are made.
2. Mission isolation. The JAUS components can typically be assembled such that a variety of missions can be supported.
3. Computer hardware independence. No assumption of or requirement of particular computer hardware is made. This allows for future adaptability and enhancement as new computer hardware becomes available.
4. Technology independence. This is similar to the computer hardware independence, but focuses more on the technical approach rather than the computer hardware. For example, many approaches could be used to determine vehicle position and orientation. No one approach, such as GPS, inertial dead reckoning, or landmark-based navigation for example, is specified.

As currently defined, JAUS Reference Architekture (RA) establishes a pre-defined set of standard, yet flexible, components that provide a menu of capabilities that can be drawn from to design an unmanned system. Components are divided into five categories:

- Command and control components
- Communications components
- Platform components
- Manipulator components
- Environmental sensor components

The RA also defines a standardized messaging construct (header and content) that enables JAUS components to exchange information in an efficient and robust fashion. The messaging approach first defines the content and usage of a standardized JAUS Header. It then prescribes the legal JAUS data types that can be incorporated into a message. Then it defines each JAUS message.

The Adaptive Planning Framework Reference Implementation was developed within such components and using such messages as defined by JAUS. Specifically, the Reference Implementation for the Adaptive Planning Framework was cast in an operational JAUS-compliant vehicle. However, the concepts and ideas that make up the Adaptive Planning Framework are not tied to nor specifically depend on JAUS. This enables other organizations to implement the framework in an alternative architecture or in future evolutions of the JAUS Reference Architecture itself.

NIST 4D/RCS

The National Institute of Standards and Technology (NIST) has been working for over two decades on establishing a standardized approach to the intelligent control of unmanned vehicle systems. The most comprehensive summary of their approach is given in NISTIR 6910, 4D/RCS: A Reference Model Architecture [27]. The 4D/RCS architecture is itself derived from NIST RCS, a domain-independent architecture developed by NIST a decade plus earlier (see [26] for a good overview of the generic RCS methodology). 4D/RCS goes on to specialize RCS to the domain of intelligent vehicle systems for military use.

4D/RCS focuses on ways to ensure that military missions involving unmanned vehicles can be analyzed, decomposed, distributed, planned, and executed in an intelligent, effective, efficient, and coordinated fashion. 4D/RCS describes in detail the functions and associated interfaces necessary to provide sensory processing, world modeling, knowledge management, cost/benefit analysis, and behavior generation. Of particular interest is its hierarchical treatment of time, providing a temporally layered set of eight planning/execution regimes (see Fig. 1). For example, it suggests that a vehicle Subsystem Planner (Level 3) ought to execute at ∼1–5 Hz with a 5 second, 50 meter planning horizon at a 40 cm grid map resolution, while a Section Planner (Level 5) might need to re-plan every 50 seconds, with a 10 minute, 5 km planning horizon at a 40 m grid map resolution.

A challenge posed by 4D/RCS is that their hierarchy of nodes calls for each node to possess a complete set of functional capabilities (i. e., World Model, Value Judgment, Behavior Generation, etc.), scaled and scoped to its level of operation in the hierarchy. The partitioning, decomposition, and distribution of the Adaptive Planning Framework Specialists and Decision Brokers across a 4D/RCS hierarchy will be a completely new research area. Of greater concern is that 4D/RCS puts a great deal of power and functionality into their World Model, including prediction and simulation.

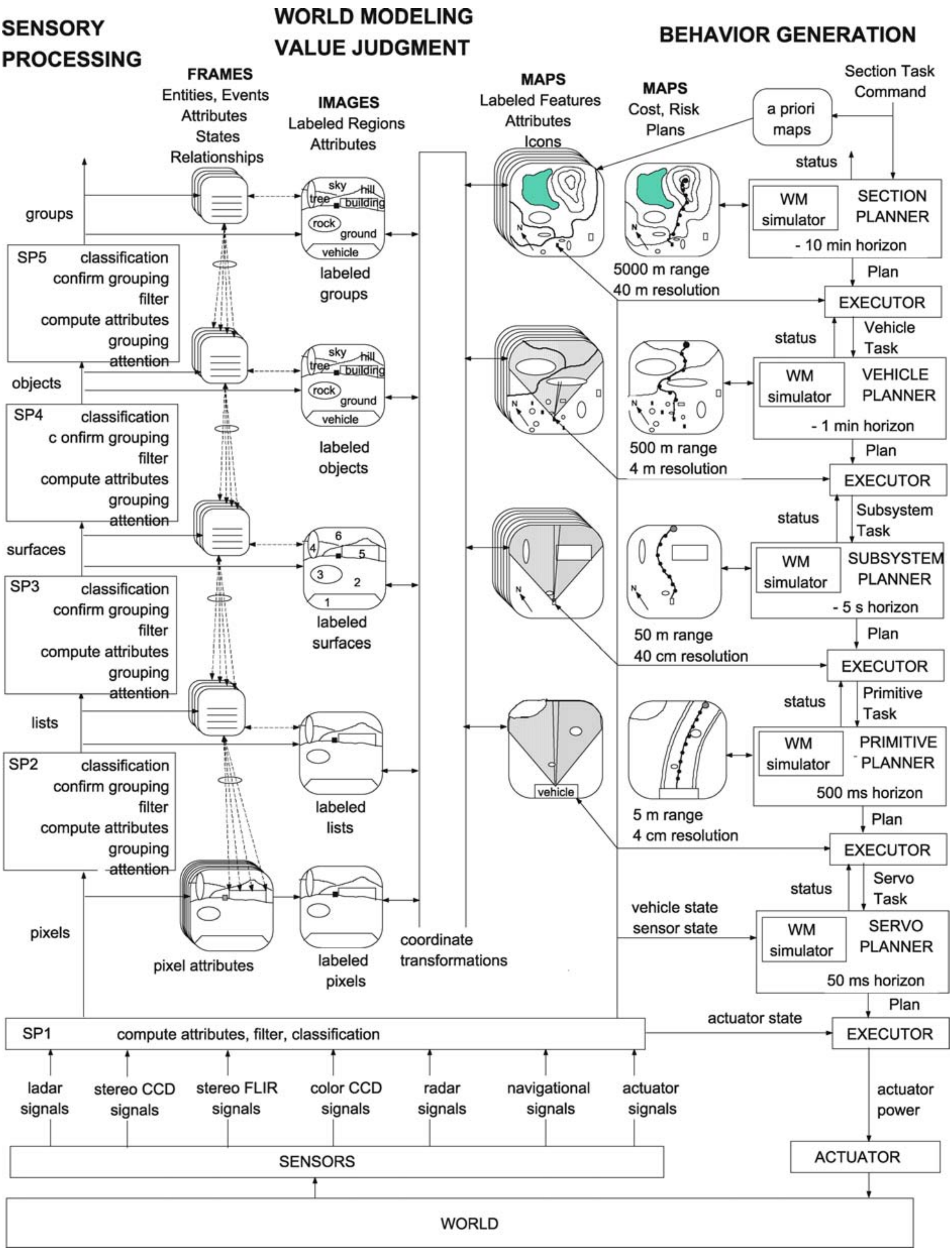
Service Oriented Architecture/ Component Oriented Architecture

Several facets of the Information Technology (IT) sector have been working to establish standards that support software interoperability across diverse organizations under the moniker of Service Oriented Architecture or SOA. SOA enables loose coupling among diverse software entities across a common network. This is accomplished by maintaining a strictly enforced standardized interface among the entities and a standardized messaging construct that enables one entity to request a service from another entity and for that service provider to send its response. This rapidly emerging standard is of interest here because the JAUS Working Group has begun a transition to a SOA-style architecture.

The most mature of these efforts is sponsored by the World Wide Web Consortium (W3C), which relies heavily upon SOA as the foundation of its Web Services initiative and, therefore, is leading the way in its maturation and adoption as a standard. Web Services extend the SOA concept to address anonymous entities that can discover one another and engage one another's services autonomously over the World Wide Web. They have published a treatise on the Web Services Architecture that includes an excellent overview of SOA in Sect. 3.1 of [45]. They go on in that section to outline some of the pitfalls of a SOA, such as network reliability and latency, lack of shared memory between service provider and consumer (i. e., everything that must be conveyed from one entity to another must be done explicitly via message content, and side effects of receipt of a message must be well understood and agreed upon), concurrency mismatches, and so on.

Industry has also taken a strong role in promoting SOA as a de facto standard. IBM (http://www-128.ibm.com/developerworks/webservices/standards/), Sun Microsystems (http://java.sun.com/developer/technicalArticles/WebServices/soa2/SOATerms.html#soaterms), and Microsoft (http://msdn.microsoft.com/architecture/soa/default.aspx?pull=/library/en-us/dnmaj/html/aj1soa.asp), to name three, have all embraced the notion.

Academia has also been active in this arena. IEEE Computer Society has formed a Technical Committee on Services Computing (http://tab.computer.org/tcsc/



Software Architectures for Autonomy, Figure 1

Excerpt from NIST PowerPoint Presentation (Source: http://www.isd.mel.nist.gov/projects/rcs/presentationhui/sld019.htm)

link.htm), and ACM has been actively including SOA topics in many of their conferences and symposiums.

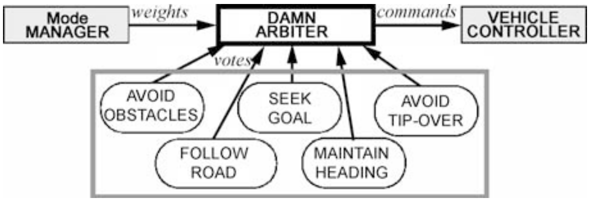
A closely related predecessor to SOA is component-based architecture (COA), which differs primarily in its stronger predisposition of what services a software entity ("component") will provide and less standardization of how components communicate with each other. In other words, COA does not worry so much about a component performing a single task (as in SOA) as long as the multiple services provided by a given component, and the interface for executing those services, are well documented. The emphasis is on providing a good platform for problem decomposition and loose coupling among components, with less emphasis on component interoperability. Aksit [1] provides an excellent compilation of articles on the topic of COA, especially Chap. 3, "Component–Based Architecting for Distributed Real-Time Systems", which, in turn, includes a detailed example of using a COA to devise a Car Navigation System (page 85). All in all, SOA can be considered a maturation, and perhaps specialization, of COA.

Distributed Architecture for Mobile Navigation

The Distributed Architecture for Mobile Navigation (DAMN) was originally published as a Ph.D. Dissertation [35] and, while not as widely adopted as the architectures discussed above, it has provided many useful insights. Even though the scope of DAMN is limited to navigation and obstacle avoidance, its distributed approach, its support of hybrid planning and implementation styles, its blend of centralized and decentralized processing, and its thoughtful treatment of salient challenges to real-time decision-making all make it worthy of elaboration here.

The basic premise behind DAMN is that centralized arbitration of distributed decision-making processes provides a reasonable and useful balance between the demands for real-time responsiveness and the challenges brought about by the asynchronous, latency-filled, heterogeneous, uncertain environment encountered by an autonomous ground vehicle. As in the other architectures discussed, DAMN provides a modular, extensible, and interoperable framework for supporting the generation and arbitration of sensor data, behaviors, and commands to the mobility platform, controllers, and actuators. This notion is shown schematically in Fig. 2, where sensor data and high-level commands have been bundled with the assortment of behaviors depicted.

The treatise goes on to present and analyze alternative structuring of the placement of arbitration (e. g., sensor vs. command vs. effect) and to explore various action selection schemes. A detailed presentation of the DAMN implementation on a CMU Navlab AGV and the experimental results achieved provides further insights into the merits and shortcomings of the architecture. Another major contribution of that research was the application of utility theory to the behavior arbitration process, as further discussed later in this article.



Software Architectures for Autonomy, Figure 2
DAMN Arbiter and Behaviors (Source: [35], page 9, Figure 1–2)

Situation Assessment

The situation assessment domain of interest to the topic of software architectures for autonomy is that of an unmanned system understanding its surroundings and status at a higher, more abstract level than that provided directly by its perception systems. In reviewing the literature, one must filter the use of the term when used in the context of the design of manned combat systems; such references often address such topics as own and enemy radars, missile tracking, and weapon lethality. Most such references are in the context of providing situational assessment for a human [16,48], such as pilot support on board a combat aircraft. Of interest here, however, is the applicability to unmanned systems, wherein the raw sensor and signal data is processed into more general situational conclusions, usually as a result of some form of inference or deduction. For clarity, the term "situation assessment" when used in this document will refer to this latter connotation. This domain is sometimes mentioned in the literature as "situational awareness" and could be referring to either of the connotations discussed above.

Work at USC [49] described the use of templates and patterns to provide situation assessment in virtual humans. They demonstrate a way to use situation assessment to improve decision-making by allowing the software system to better focus its attention (i. e., computing resources) with the goal of improved utilization of onboard resources.

Of particular interest is the work underway at NIST. They are working in several areas that address situation assessment. One has to do with incorporating situation assessment feedback to human operators of robotic devices [40]. While their emphasis is on the human-machine interface, there are insights to be gained from the situation identification and classification schemes that they developed. An even more relevant front is their work on using 4D/RCS to control on-road robotic vehicles. There are both formal papers [38] and materials and presentations available on the NIST web site (see Fig. 3) that demonstrate ways to incorporate situation assessment notions into the 4D/RCS architecture.

Weiss, Philipps and To et al. [46] present a capability that could be adapted into a Traffic Specialist. It provides situation classification and prediction for an assortment of expressway-related conditions (such as same/different lane assignment for other vehicles that are detected), and states (Such as approach rate {approach | approach with distance warning | approach with collision warning}). Similarly, [13] introduce material that could form a Collision Avoidance Specialist that could manage interactions with moving obstacles using such notions as "time to collision", "time to brake", and "time to disappear". Another area of interest is vehicle self-awareness and work such as [33] sheds light on how a Vehicle Health Specialist might be devised.

Finally, it should be noted that much of the discussion of situation assessment in the literature was secondary to a broader discussion and is, thus, of most use in providing insights into possible nomenclature and classification. References such as these are discussed in the Knowledge Representation section rather than here.

Planning and Decision Making

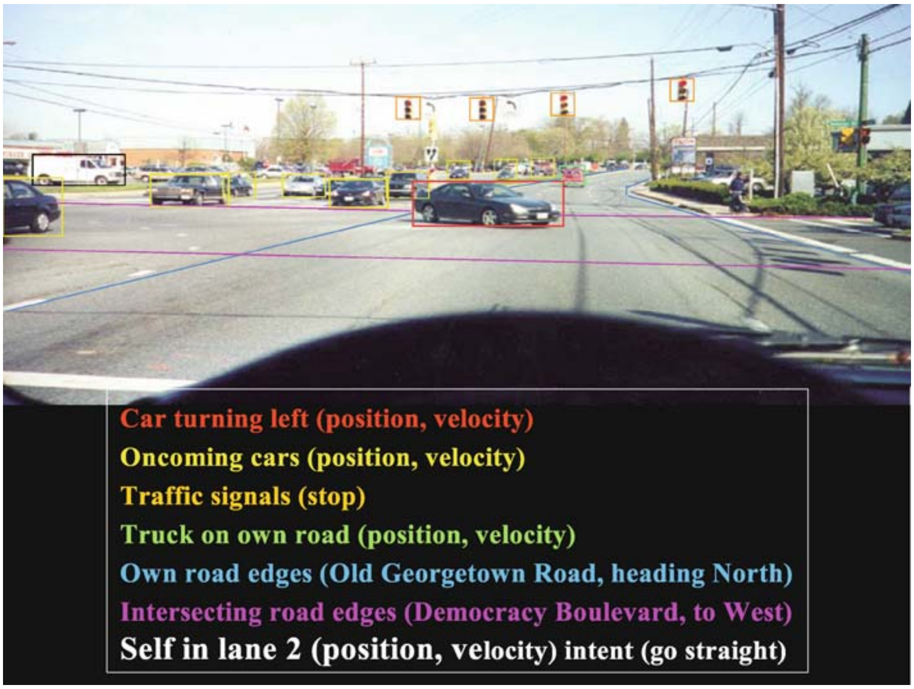
Since the scope of this topic is so broad, its treatment here will be, first of all, limited to the domain of real-time planning and decision-making on an AGV and then further organized as an assortment of "views". The notion of planning refers to the orchestration of executable behaviors to achieve a goal (e. g., find a series of waypoints that will take the vehicle to a desired goal, then drive the vehicle to those waypoints while avoiding obstacles, obeying driving rules and maintaining stability), as well as the low-level planning conducted by a given behavior (e. g., finding an obstacle-free path towards the next waypoint within the perception horizon of the vehicle). The following list of behavior primitives are but a sampling of those gleaned from the literature:

- Seek Goal
- Avoid Obstacles
- Follow Road
- Respond to Blocked Path
- Explore
- Wander
- Maintain Stability
- Seek Target/Intruder
- Intercept Target
- Mark Location

Viewed as a Sense-Plan-Act Problem

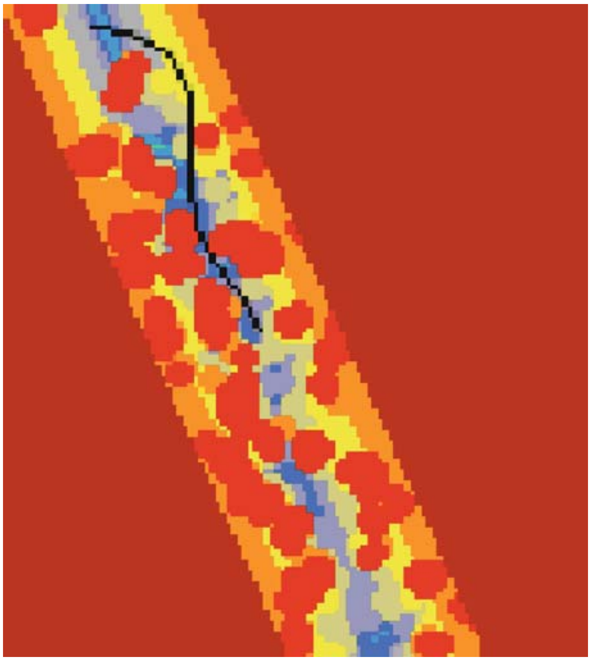
This is perhaps the most fundamental view of autonomous control of a mobile robot and one into which many autonomous robotic implementations can be cast. The notion is to neutralize uncertainties in the robot's perception of its world, its understanding of its own state, and the effects of its own actions by indirectly "closing the loop" through the continuous gathering of feedback from its environment while executing its plan [25]. Since it is anticipated that the plan itself will be divided into a sequence of steps, the idea is that the results of executing the initial steps can be observed and compared with expected results. If expectations are not being adequately met (in essence, forming an "error" signal), then the subsequent steps can be adjusted accordingly, or an entirely new plan can be published. It is presumed that the robot will have an ability to store its perception and state knowledge in some form of a world model, which can, in turn, be used by the planner.

This design style best describes the autonomous control used on the NAVIGATOR [8]. The four environmental sensors publish their findings, in the form of a traversability grid, to a sensor arbiter. Two additional pseudo-sensors each publish a traversability grid to the sensor arbiter denoting the a priori route boundary and a priori path plan. The sensor arbiter then fuses these inputs and publishes to the receding horizon planner a comprehensive traversability grid, which represents a localized view of a world model. The receding horizon planner uses an A* search algorithm and a simple vehicle model to iteratively produce viable plans that achieve a desired goal state and to choose the one that minimizes traversability cost. The goal itself is based on the a priori path plan and is replaced with a new goal once the vehicle nears it. The planning search that occurs has as its only objective the publishing of an instantaneous wrench command (steering, throttle, brake) to the vehicle's primitive driver, whose job is to execute that wrench as actuator positions. Thus, every cycle of the planner produces a new wrench command. Since every component in the chain executes at a nominal rate of 20 Hz, a new "plan" (as manifested in the instantaneous wrench command) is always being issued, thus providing a responsive behavior, with some deliberation on how that behavior is generated. Figure 4 shows a snapshot of an arbitrated traversability grid and the instantaneous plan. Red, orange, and yellow indicate lessen-



Car turning left (position, velocity)
Oncoming cars (position, velocity)
Traffic signals (stop)
Truck on own road (position, velocity)
Own road edges (Old Georgetown Road, heading North)
Intersecting road edges (Democracy Boulevard, to West)
Self in lane 2 (position, velocity) intent (go straight)

Software Architectures for Autonomy, Figure 3
Excerpt from NIST PowerPoint Presentation (Source: http://www.isd.mel.nist.gov/projects/rcs/presentationhui/sld061.htm)



Software Architectures for Autonomy, Figure 4
Example Traversability Grid taken from the NAVIGATOR while in a Cluttered Roadway

ing severity of obstacles, gray and blue indicate improving degrees of smoothness of terrain. The instantaneous plan is indicated in black.

One difference in the NAVIGATOR's implementation of the Sense-Plan-Act paradigm is that, by encapsulating the a priori plan into a pseudo-sensor whose findings compete with those of the other sensors, the conventional aspects of planning provide only "suggestions" for a preferred action, rather than forcing the vehicle onto a defined course. Although implemented quite differently, this notion is in concert with the findings of Payton, Rosenblatt and Keirsey [29], who go on to note that "In general, internalized plans should be conceived as representations that allow the raw results of search in an abstract state space to be made available as advice to continuous real-time decision-making processes".

There are many good examples of robotic systems that have implemented some fashion of the Sense-Plan-Act paradigm. Most have to do with navigation and obstacle avoidance, such as Batavia and Nourbakhsh [6]. Examples of this paradigm applied to other aspects of robotic planning and decision-making are much harder to find.

Viewed as a Subsumption Problem

The notion of empowering a mobile robot to operate without any centralized control was first introduced by Rodney Brooks as he devised a self-managing, layered control scheme dubbed the "subsumption architecture" [7]. By decomposing a robot's control system into layers of task-achieving behaviors, control is governed by the dominant layer in play at an instance in time, which, in turn, "subsumes" the behaviors of the lower level layers. For example, let "Wander" be considered a level 1 behavior and "Explore" a level 2 behavior. Since Explore is the higher level, it will self-determine whether exploring is an appropriate behavior under current circumstances. If so, then it will alter the Wander behavior to be not random, but to fulfill its wishes to visit new areas. If not, then it will allow the Wander behavior to proceed without any alteration. This notion is extrapolated across all possible behaviors. This style of planning and decision-making is often referred to as "reactive". The resultant behavior of the robot is referred to as "emergent" since it is likely that the observed behavior is some extemporaneous blend of the possible behaviors that the robot could execute.

This approach to planning and decision-making has a dedicated following and is especially appealing for multi-agent and swarm applications. For example, the subsumption architecture and reactive behavior play a major (though not exclusive) role in the design of robots at the Idaho National Lab (see www.inl.gov/adaptiverobotics).

The differences between these first two views can be captured by the relative importance placed on each of the three components of the Sense-Plan-Act view. For example, a purely reactive system does almost no local planning since every stimulus anticipated during the sensing stage has a prescribed behavioral action, thus relegating the planning stage to simply resolving action conflicts when more than one stimulus is perceived or queuing and dispensing actions when one stimulus invokes multiple actions (i. e., managing the subsumption process). Conversely, a deliberative system will have a large emphasis on the planning stage, attempting to formulate a new plan that incorporates newly sensed information along with any changes in state of the vehicle or its mission while simulating the effect of alternative actions on the quality and viability of the plan. The juxtaposition of the Sense-Plan-Act view's emphasis on deep planning through possibly time-consuming deliberation and the Subsumption view's potentially unpredictable, but fast, reaction to stimuli, explains why researchers are still seeking other, hybrid or blended, planning and decision-making styles.

Viewed as a Decision Theory Problem

Another rich area of exploration is how classical decision theory might be applied to the AGV domain. For example, Karacapilidis and Papadias [20] describe how argumentation can be automated and used to support collaborative decision-making. Their ideas for automating argumentation constructs include evaluations such as "Scintilla of Evidence", "Beyond Reasonable Doubt", and "Preponderance of Evidence".

Rauenbusch and Grosz [32] and others speak of devising explicit "Plan Trees" whose nodes encapsulate the desired action/behavior, associated constraints, and contextual applicability and whose structure models the desired decision-making outcomes. The search through the tree is conducted using some measure of cost or value such that the correct path through the tree delivers the correct series of actions/behaviors.

Hoffman and Yates [15] present a synopsis of what has become known as the "three-step decision-making process". In this paper, they report that most, if not all decisions can be modeled as a cascading set of three-step activities. One of the models specifically referenced for use in process control is 1) Situation Assessment, 2) Planning, and 3) Commitment to a course of action. Each of these steps may be expanded into another three-step decision-making process, such as deciding which situational conditions are present or relevant, or whether to keep or abandon a committed action.

A final realm under decision theory is known as Hierarchical Task Network (HTN) planning [9] provide an overview of this concept and cite the seminal works that have contributed to it on the way to introducing a formalism of HTN planning semantics. The basic premise of HTN planning is to iteratively decompose tasks until primitive tasks are reached (defined as tasks that cannot be further decomposed and that are actionable). These primitive tasks are assembled into a network of increasingly abstract tasks allowing a planning algorithm to select a high-level task, recursively expand its children until its primitive tasks are reached. Some expansions may be constrained based on the current situation, thus pruning the search when compared with an unconstrained expansion of the network. Each reachable path from the high-level tasks to the primitive tasks becomes a candidate plan. While this exploration of the HTN is taking place, the candidates are being evaluated by so-called "critics" so that any arising conflicts can be identified and the winning candidate declared. Because of its deep reasoning, HTN-based planning is not typically used for real-time applications.

Viewed as a Behavior Arbitration Problem

The concept of Behavior Arbitration was introduced as part of the Distributed Architecture for Mobile Navigation

(DAMN) [35] as a key ingredient for achieving its goal of balancing centralized and decentralized design styles. All (decentralized) behavior generators submit their control output (referred to as a "vote") and the (centralized) DAMN Arbiter fuses their votes into a single command set to the vehicle.

"Utility fusion", which uses traditional utility theory to provide an alternative to command fusion, is another concept that evolved from DAMN [36]. This notion requires each behavior generator to submit a probabilistic utility estimate along with its vote, thus enabling a "utility arbiter" to compute the Maximum Expected Utility and use it to select the optimal behavior.

Viewed as an Action Selection Problem

Action Selection is another way to view planning and decision-making on an AGV. In this view, the mobile robot is tasked with selecting the most appropriate action based on the current situation, in spite of inaccurate, incomplete, and possibly unforeseen information. For this to happen, there must exist some catalog or repository of possible actions from which to select and the criteria upon which to base a selection decision. Pirjanian [30] provides an excellent overview of ten varying approaches to the action selection problem. In this treatise, he summarizes each (including DAMN), then compares and contrasts them in terms of eight criteria, including planning vs. reactivity, synchronous vs. asynchronous, hierarchy vs. no hierarchy, and knowledge representation which all have a direct bearing on the current research.

NIST has also developed an approach to action selection via its hierarchical planning and control scheme [21, 22]. The scheme enables the system to plan at different rates at each level, with the scope of planning fixed for each level. For example, high-level goal planning might take place at a lower resolution and update rate, but would cover a larger expanse than say planning for obstacle avoidance. The plans themselves are broken into a tree or graph of subgoals and subtasks (task decomposition itself is discussed under Knowledge Representation) and the actions are selected, executed, and monitored in accordance with the defined planning levels. The planning levels are chosen to be consistent with the time, duty cycle, and range horizon parameters established in the 4D/RCS architecture. For example, AGV mobility planning is broken into four levels: Servo, Prim(itive), Autonomous Mobility, and Vehicle System. Balakirsky and Lacaze [2] elaborate how planning, in the form of Value Judgment and Behavior Generation, takes place for the Vehicle System planning level.

Viewed as an Adaptive Planning Problem

Note that in some literature, "adaptive" is used to mean that the system "learns" from its experience, thus improving its performance over time, whereas the connotation used here is that the system alters its plan based on new, situational information that has been provided by upstream knowledge and data processing. Thus, while the possibility of actually changing the a priori behaviors from which to choose through learning should not be ruled out for future generations of the Adaptive Planning Framework, it is certainly not the emphasis or the motivation for using the term "adaptive" in its moniker. The genesis of adaptive planning as used here was a search to improve the performance of (manual) military mission planning through the use of expert systems, such as the Adaptive Mission Planning System in [41]. The quest continues as military planners seek to reduce 24-month planning cycles down to a year or less for complex deployments and even less for Crisis Action Planning [14]. In fact, their definition, "Adaptive Planning is the systematic, on-demand creation and revision of executable plans, with up-to-date options, as circumstances require", could suffice for the work conducted here as long as its transition to an autonomous, real-time setting is understood.

The need to alter a plan already in progress can have a number of causes, including insufficient time for completion, ineffective results, changes in the situation, and receipt of a new objective to name a few. The Artificial Intelligence community has driven related work in this area, but application to mobile autonomous robotics has not been at the forefront. For example, [12] presents an excellent treatise of an adaptive planning architecture based on the premise that an "agent dynamically constructs explicit control plans to guide its choices among situation-triggered behaviors". To accomplish this, she identified and explored five areas where an intelligent system might require adaptive behavior, depending on the situation encountered:

1. Perception Strategy – Adapt to information requirements and resource limitations
2. Control Mode – Adapt to goal-based constraints and environmental uncertainty
3. Reasoning Tasks – Adapt to perceived and inferred conditions
4. Reasoning Methods – Adapt to available information and current performance criteria
5. Meta-Control Strategy – Adapt to dynamic configurations of demands and opportunities

As an example of more recent work that does focus on mobile robotics in a real-time setting, Hassan, Simo and Crespo [11] offer a behavior-based architecture that will adapt to temporal constraints by allowing itself to utilize more deliberative techniques when time is available, but moving towards more reactive behaviors when time is at a premium. They also introduce the notion of adjusting the quality of service that a given element might deliver based on the situation encountered. For example, this approach might allow the system to attempt to achieve its goal with a "rough" plan if a "complete" plan could not be delivered in a timely enough manner. Musliner [23] and his Adaptive Mission Planner, provides another view on how to empower an autonomous system to alter its plans based on temporal constraints and in light of changing environments, objectives, and system capabilities. That work built upon his earlier efforts to devise the Cooperative Intelligent Real-time Control Architecture (CIRCA) [24], which provides formalisms on how to represent tasks and decisions in a LISP setting. While CIRCA has not been applied in the mobile robotic domain (making its suitability to support an AGV unknown), there are insights to be gained from this work. Finally, NIST has incorporated an element of adaptive planning in their recent work on autonomous on-road driving as part of 4D/RCS. For example, Balakirsky and Scrapper [3] discuss an expert system and knowledge representation scheme that support adaptive planning for autonomous lane and speed management.

Knowledge Representation

In this section, related work on Knowledge Representation relevant to the domain of AGVs is explored. Knowledge Representation refers to the schemas and constructs used to document, standardize, normalize, and utilize the entities within the domain of interest. It must capture the semantics and meanings of the relationships among the entities, as well as their names, descriptions, attributes, and the method or reasoning mechanism for determining their current state or value.

Sources of such domain knowledge include technical documents, specifications, training manuals, etc. (many of which can be accessed via the web). Example knowledge sources include a table of Autonomous Mobility Situation Coverage Requirements from Demo III requirements analysis [34], a Functional Taxonomy chart for an AGV from a TACOM (the U. S. Army's Tank-Automotive COMmand) PowerPoint presentation [31], and by drawing analogies from human military operations as found in the Army Universal Task List [44]. Remaining knowledge gaps must be filled in by interviews of subject matter experts or perhaps empirically through experimentation.

By far, the most work in knowledge representation for intelligent vehicles has been done by NIST. Thus, this section will conclude with an extended example, demonstrating their approach to representing knowledge about situational conditions, states and events, planning, and behaviors within the 4D/RCS context.

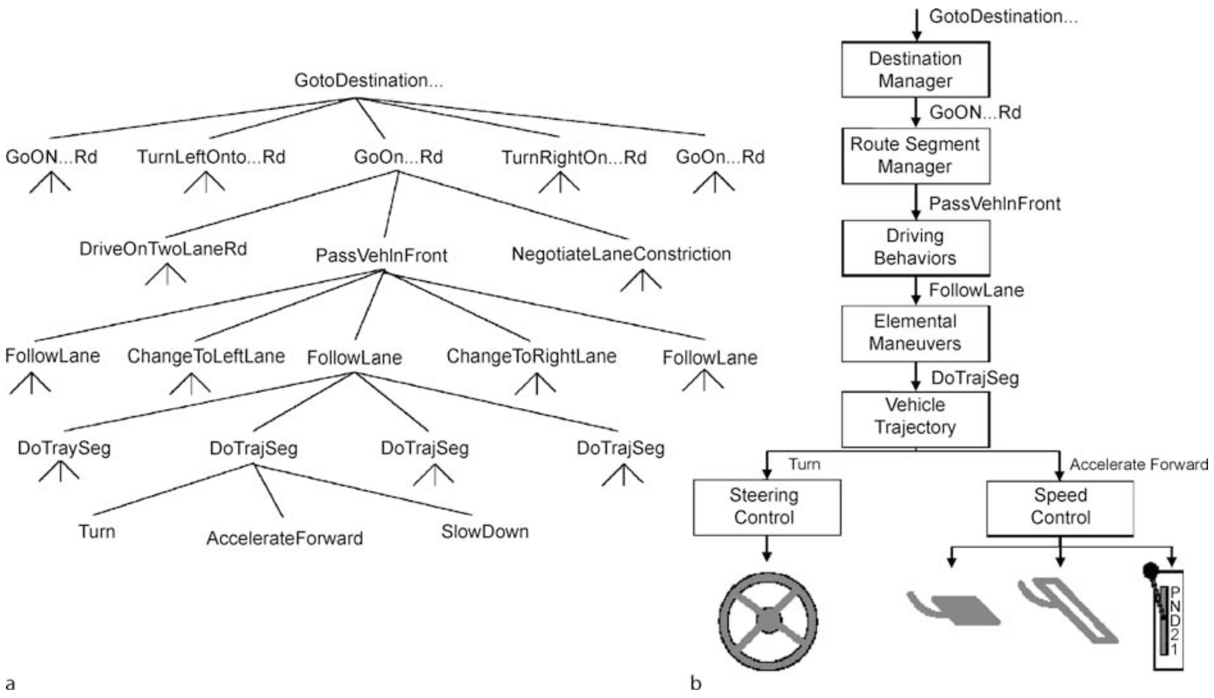
Lexicons, Taxonomies, and Ontologies

One technique for knowledge representation is to progress from a lexicon (a domain-specific dictionary of terms), to a taxonomy (a logical ordering and categorization of those terms), to an ontology (an explicit specification of those terms along with the semantics and relationships among them). One on-line dictionary defines ontology as follows:

> An explicit formal specification of how to represent the objects, concepts and other entities that are assumed to exist in some area of interest and the relationships that hold among them ... Definitions associate the names of entities in the universe of discourse (e. g. classes, relations, functions or other objects) with human-readable text describing what the names mean and formal axioms that constrain the interpretation and well-formed use of these terms ... The hierarchical structuring of knowledge about things by subcategorizing them according to their essential (or at least relevant and/or cognitive) qualities [17].

Much work is in progress attempting to build general purpose, or even "common sense" ontologies that would be useful to all domains. The most famous of these is the OpenCyc project (http://www.opencyc.org/), which, with 47,000 concepts and 306,000 assertions about them to date, is well on its way to achieving its vision to become "the world's largest and most complete general knowledge base and commonsense reasoning engine". Another initiative of interest is the DARPA Agent Markup Language (DAML) project (http://www.daml.org/), which was sponsored by DARPA to create an xml extension that provides (among other things) a rich and suitable language for the creation of general-purpose ontologies (282 distinct ontologies had been created using DAML by the time the program funding was terminated in 2006 and the work absorbed by W3C).

For the AGV domain, the scope of the knowledge that must be represented via the techniques discussed in this subsection is still quite broad. Situational knowledge spans from urban to highway to off-road environments,



Software Architectures for Autonomy, Figure 5
NIST Knowledge representation schemes for on-road driving. (Source: [4,5], Figures 3 and 4)

potential obstacles and hazards that might be present, traffic rules and driving best practices, and so on. Planning and behavior knowledge encompasses a wide variety of missions and tasks, whether they are high-level (conduct search and rescue operation), tactical (pass the vehicle), elementary (change lane), behaviors (avoid obstacles, maintain stability), planning rules and processes, and so on. Even knowledge about "self" or "ego" must be represented, such as capabilities, limitations and constraints, or current status. There is a major initiative under way at NIST, sponsored by TACOM, to develop an Intelligent Systems Ontology. Although still a work in progress, this intelligent-vehicle-specific ontology is expected to provide a standard set of domain concepts, their attributes and their interrelationships, delivered in a fashion that facilitates knowledge capture and reuse [39]. This ontology is beginning to gain traction as it makes its way into the AGV navigation planning community outside of NIST [37].

World Model Knowledge Store (WMKS)

Another dimension of knowledge representation is how data, information, and knowledge are stored. Whether it is provided a priori, or it is perceived, inferred, or received by the AGV, there must be a place and a format for storing, accessing, and analyzing it. Such data, information, and knowledge are often referred to as the "world model" and the place where they are stored as the "knowledge store". The breadth and sophistication of the world model knowledge store for a given AGV design will vary widely, depending on its degree of autonomy, the scope of its behavior, the complexity of its design, etc.

Situation Assessment findings, which must also be managed, fit into what some communities refer to as "meta-knowledge", i. e., knowledge about the knowledge. For example, while pumping out its perception data, a sensor could independently assess and report on its own confidence in its findings and its own health, and perhaps even declare that its own results should not be used right now (say, due to a camera white-out).

Although not always so, the knowledge store is usually persistent, using either a relational database or an object-oriented knowledge based system. Since much of the information stored is of a geo-spatial nature, the knowledge store often includes geo-spatial extensions for explicitly representing GIS and topographical data, polygonal objects, etc. Another consideration is whether the WMKS contents are stored in a central location, accessible by all AGV modules (sometimes referred to as a "blackboard architecture") or each module maintains a subset of the WMKS containing just the content it needs, with data, information, and knowledge marshaled among the AGV modules on an as-needed/as-requested basis (sometimes referred to as a "publish/subscribe architecture").

Knowledge Representation at NIST

NIST advocates task decomposition as a key knowledge representation technique to support the hierarchical control strategy emphasized in its 4D/RCS architecture and has published widely on various ways to accomplish it [4,5]. This technique for representing the actionable elements that could be assembled to create a plan strives to break high-level tasks (e. g., a mission objective) into distinct hierarchical levels and also to identify multiple subtasks at a given level. Figure 5a–5b shows an example of how the "GoToDestination" task is decomposed into a "planning graph" that ultimately leads to a specific wrench command to the vehicle. The system must know (or be able to infer) the state of each node in the tree along with the cost of each arc in order for the associated control module to formulate the appropriate plan. Extending the example in Fig. 2–5b, a Destination Manager has determined that staying on the current road is appropriate and a Route Segment Manager has decided that passing the vehicle in front of it is the most desirable way to reach the destination. A Driving Behaviors module knows that its own vehicle has already changed into the passing lane and has further determined that the best thing to do right now is to stay in that lane, while a low-level Elemental Maneuvers module has found a wrench that ought to produce the requested outcome. Each Manager or module manages its own situational understanding either from direct sensory input or from its own local subset of the World Model Knowledge Store. Naturally, there are other tree elements and control modules that address following distance, speed, and so on, in addition to non-mobility-related tasks, such as payload management, communications, etc.

Once a plan is devised and approved, its elements must be executed by invoking one or more actions or behaviors, or perhaps by unleashing an entire subsystem to take over low-level control of the vehicle. NIST advocates the use of State Tables to represent the action decision-making knowledge [4]. A State Table is crafted for each node in the Task Decomposition Tree containing the rules that the control module is to use for mapping node inputs (states or situations) to allowable output actions.

To trigger the appropriate and desired state response, the matching situation must be known. The NIST approach to this is to determine and store the cascading precursor situational knowledge as a collection of "world states", but, in conformance to the 4D/RCS architecture, only that subset relevant to a given module. The lane-changing example concludes with a glimpse of the dozens of situational findings that lead up to the finding of interest ("ConditionsGoodToPass").

Adaptive Planning Framework

The Adaptive Planning Framework was developed by the authors at the University of Florida in order to address the requirements of the DARPA Urban Challenge that was held in November 2007. Figure 6 shows the University of Florida Navigator vehicle on which the Adaptive Planning Framework was implemented for the DARPA Urban Challenge. In the Adaptive Planning Framework, the system is assumed to be able to operate in a finite number of behavior modes. These behavior modes govern how the vehicle operates under various driving conditions. The framework is predominantly used to make intelligent decisions pertaining to these behaviors. The framework is scalable to systems of varying complexity and size and is compatible with existing architectures such as JAUS RA-3.2, NIST 4D/RCS, and others. The Adaptive Planning Framework is composed of three principle elements tasked with assessing the situation, determining the suitability and viability of all possible solutions, and executing the most suitable of all recommended solutions.

Situation Assessment Specialist

Dynamic environment information, originating from any array of sensors is monitored and managed by the Situation Assessment Specialists. Each specialist design is tailored to the sensor or collection of sensors whose data it analyzes. These specialists can, but are not required to, "live" on the same computing node that directly receives the sensor input. While the inputs to the specialist can come from any data source, the output or "finding" must adhere to specific guidelines outlined by the framework. Findings can be in the form of conditions, state, or events. A condition may have a value of present or absent only. All conditions are by default absent and must be proven present at each iteration. A finding classified as a state can only exhibit one of many a priori states. The event category is reserved for findings whose occurrence at some point in time is of significance even after the initial finding has passed. Once the findings have been generated the information is disseminated to all other components that might need it.

An example of a situation assessment specialist would be a software component whose sole function was to determine if it is safe to move to the adjacent lane. This component would monitor sensor data and reach a Boolean

Software Architectures for Autonomy, Figure 6
Team Gator Nation NaviGATOR

conclusion that would be stored as metadata for use by other processes. A second example would be a software component whose sole function was to determine if it is 'legal' to move to an adjacent lane. Here 'legal' is defined as not crossing a yellow line or not changing lanes when approaching an intersection.

Behavior Specialist

The findings rendered by the Situation Assessment Specialists are consumed by the behavior specialists. There is a one-to-one mapping of each behavior with a behavior specialist. The role of the specialist is to monitor the findings and evaluate the suitability of its behavior under the current perceived operating conditions. As with the specialist findings, the default recommendation is unsuitable and must be proven appropriate at every iteration of the program to ensure truth of the results and operating safety. A specialist does not possess the ability to activate or deactivate its associated behavior; such authority is only given to the Decision Broker.

For the DARPA Urban Challenge problem, the vehicle was programmed with six behavior modes. The corresponding behavior specialist constantly evaluates the appropriateness of its behavior. The six behavior modes were:

1. Roadway Navigation. The Roadway Navigation behavior is the primary driving behavior deriving commands to be sent to the vehicle actuators while the objective is lane following. This behavior allows the vehicle to navigate the roadway within the lines of its desired lane and maintain a safe following distance behind any vehicles ahead.
2. Change Lane Maneuver. The change lane maneuver is used in passing situations or in cases where the vehicle must change lanes in a multi-lane road in order to pass through a mission goal point. The behavior constrains the vehicle to remain within the lane boundaries of the new lane.
3. Reverse Direction. This behavior is called whenever it is determined that the current lane is blocked and there is no alternate clear lane available for passing. It is also applicable in cases where the vehicle has entered a 'dead end' road that it must "escape" to reach a mission goal point.
4. Intersection Traversal. The intersection traversal behavior is applicable when the vehicle enters the vicinity of an intersection. This is one of the most complicated behavior modes in that the system must rely on a series of situation assessment specialists to safely navigate the intersection. This behavior mode must handle queuing, stopping at the stop line, determining right of way, and ultimately traveling through the intersection while avoiding other vehicles.
5. Open Area Navigation. Open area navigation is a behavior that is only needed in special circumstances. This behavior allows the vehicle to move towards a goal location without striking any object, while avoiding any

rough terrain. This is in effect the only behavior mode that was required in the 2005 DARPA Urban Challenge. It is useful in the Urban Challenge when the vehicle is in an open area such as a parking lot prior to performing an actual parking maneuver.

6. Parking Lot. This behavior must deal with the problems that arise in the parking lot scenario where precise motion is necessary. When the vehicle approaches the vicinity of an assigned parking space, precise path planning will be initiated to align the vehicle as required. Situation assessment specialists will be monitoring the near surroundings of the vehicle to center the vehicle in its parking space while avoiding any static or dynamic objects.

Decision Broker

At the highest level of the framework lies the Decision Broker. Its role is to monitor all Behavior Specialist recommendations. It assumes ultimate authority over how the vehicle will operate while in autonomous mode. Like the other entities within the framework, the Decision Broker can base its conclusions on not only the recommendations and findings of other specialists, but it may also look at data from any other pertinent source. The author's implementation of the Adaptive Planning Framework centralized all the Decision Broker functionality within the JAUS Subsystem Commander component. The framework architecture employs an asynchronous, iterative, forward chaining reasoning approach to decision making.

Future Directions

JAUS is migrating to the Society of Automotive Engineering (SAE) Standard AS-4, Unmanned Systems. The transition to the SAE offered opportunity to further advance the standard through the introduction of more formalized specifications and methodologies.

The AS-4A Architecture Framework subcommittee has prepared document AIR 5665, Architecture Framework for Unmanned Systems. This document describes the concepts, capabilities and interoperability concerns of unmanned systems, and lays the foundation for all subsequent Aerospace Standards to be released by AS-4.

The AS-4B Network Environment subcommittee is preparing the document AIR5645, "JAUS Transport Considerations". This document examines aspects of JAUS, aspects of communications media, and aspects of the domain of unmanned systems, explaining how these affect the design of transport mechanisms for JAUS messaging. The committee has also prepared an initial draft version of AS5669, "JAUS Transport Specification".

The AS-4C Information Modeling and Definition subcommittee has prepared a working draft of the document AS5684, JAUS Service Interface Definition Language. This specification allows for the unambiguous definition of JAUS services. AS5684 is specified using Relax NG Compact notation, a machine-readable language, allowing for the creation of various tools to aid in the unmanned system design process. The AS-4C Subcommittee has also begun work on AS5710, JAUS Service Set. AS5710 is a collection of standard unmanned system services. Each service is defined using the JAUS Service Interface Definition Language. Current work on this document includes defining a discovery service and transport service, and transitioning the JAUS Reference Architecture 3.2 components to services.

Bibliography

1. Aksit M (2002) Software architectures and component technology. Kluwer, Boston
2. Balakirsky S, Lacaze A (2000) World modeling and behavior generation for autonomous ground vehicle. In: Proceedings of the 2000 IEEE International Conference on Robotics & Automation. San Francisco, pp 1201–1206
3. Balakirsky S, Scrapper C (2004) Knowledge representation and planning for on-road driving. Robotics Auton Syst 49(2004):57–66
4. Barbera A, Albus J, Messina E, Schlenoff C, Horst J (2004) How task analysis can be used to derive and organize the knowledge for the control of autonomous vehicles. In: Proceedings of the 2004 AAAI Spring Symposium Series on Knowledge Representation and Ontology for Autonomous Systems. Palo Alto, pp 67–78
5. Barbera A, Messina E, Huang H-M, Schlenoff C, Balakirsky S (2004) Software engineering for intelligent control systems. Künstliche Intell 3(4):22–26
6. Batavia PH, Nourbakhsh I (2000) Path planning for the cye personal robot. In: Proceedings of International Conference on Intelligent Robots and Systems, 2000 (IROS 2000). Takamatsu, Japan, 15–20
7. Brooks RA (1986) A robust layered control system for a mobile robot. IEEE J Robotics Autom 2(1):14–23
8. Crane CD, Armstrong DG, Touchton R, Galluzzo T, Solanki S, Lee J, Kent D, Ahmed M, Montane R, Ridgeway S, Velat S, Garcia G, Griffis M, Gray S, Washburn J, Routson G (2006) Team CIMAR's NaviGATOR: An unmanned ground vehicle for application to the 2005 DARPA grand challenge. J Field Robotics 23(8):599–623
9. Erol K, Hendler J, Nau D (1994) Semantics for hierarchical tasknetwork planning. CS-TR-3239, UMIACS-TR-94-31, ISR-TR-95-9, University of Maryland, Institute for Advanced Computer Studies College Park, MD
10. Franklin S, Graesser A (1996) Is it an agent, or just a program?: A taxonomy for autonomous agents. In: Third International

Workshop on Agent Theories, Architectures, and Languages. Budapest, pp 21–35

11. Hassan H, Simo J, Crespo A (2001) Flexible real-time mobile robotic architecture based on behavioural models. Eng Appl Artif Intell 14(5):685–702
12. Hayes-Roth B (1995) An architecture for adaptive intelligent systems. Artif Intell 72:329–365
13. Hillenbrand J, Kroschel K, Schmid V (2005) Situation assessment algorithm for a collision prevention assistant. In: Intelligent Vehicle Symposium, 2005. Las Vegas, pp 459–465
14. Hoffman H (2004) Adaptive planning overview. Presented at Military Operations Research Society, Capabilities-Based Planning: The Road Ahead. The Military Operations Research Society (http://www.mors.org/), Alexandria
15. Hoffman RR, Yates JF (2005) Decision(?)Making(?). IEEE Intell Syst 20(4):76–83
16. Howard C, Stumptner M (2005) Probabilistic reasoning techniques for situation assessments. In: Third International Conference on Information Technology and Applications, 2005 (ICITA '05). Sydney, pp 383–386
17. Howe D (2005) The Free Online Dictionary of Computing. http://foldoc.org/foldoc.cgi?query=ontology. Accessed Oct 2006
18. JAUS (2005) JAUS Reference Architecture, version 3.2. JAUS Working Group, (http://www.jauswg.org/)
19. JAUS-OPC (2005) OPC 2.75 Interface Control Document, version 1.0, Payload Interface section. JAUS Working Group – OCU and Payloads Committee
20. Karacapilidis N, Papadias D (2001) Computer supported argumentation and collaborative decision making: the HERMES system. Inf Syst 26(4):259–277
21. Lacaze A (2002) Hierarchical planning algorithms. Presented at SPIE 16th Annual International Symposium on Aerospace/Defense Sensing, Simulation and Controls, Orlando
22. Murphy K, Abrams M, Balakirsky S, Coombs D, Hong T, Legowik S, Chang T, Lacaze A (2000) Intelligent control for unmanned vehicles. Presented at World Automation Conference (WAC 2000), Maui
23. Musliner DJ (2001) Imposing real-time constraints on self-adaptive controller synthesis. Lect Notes Comput Sci 1936:143–160
24. Musliner DJ, Durfee EH, Shin KG (1995) World modeling for the dynamic construction of real-time control plans. Artif Intell 74(1):83–127
25. Nilsson NJ (1998) Artificial intelligence: a new synthesis. Morgan Kaufmann, San Francisco
26. NIST (National Institute of Standards and Technology) (1992) A real-time control system methodology for developing intelligent control systems. NISTIR 4936. National Institute of Standards and Technology, Gaithersburg
27. NIST (National Institute of Standards and Technology) (2002) 4D/RCS: A Reference Model Architecture for Unmanned Vehicle System, Version 2.0. NISTIR 6910. National Institute of Standards and Technology, Gaithersburg
28. Panzarasa P, Jennings NR, Norman TJ (2002) Formalising collaborative decision-making and practical reasoning in multiagent systems. J Logic Comput 12(1):55–117
29. Payton DW, Rosenblatt JK, Keirsey DM (1990) Plan guided reaction. IEEE Trans Syst, Man, Cybernetics 20(6):1370–1382
30. Pirjanian P (1997) An Overview of System Architecture for Action Selection in Mobile Robotics. Laboratory of Image Analysis, Aalborg University, Aalborg
31. Pritchett W (2002) A domain framework for intelligent ground combat vehicle systems. Presented at Software Technology Conference (STC 2002). TACOM (Tank-automotive & Armaments COMmand), Salt Lake City
32. Rauenbusch TW, Grosz BJ (2003) A decision making procedure for collaborative planning. Presented at AAMAS'03 – Second International Joint Conference on Autonomous Agents and Multi-Agent Systems, Melbourne
33. Reichard KM, Crow EC (2005) Intelligent self-situational awareness for unmanned and robotic platforms. AUVSI Unmanned Systems Conference, Baltimore
34. Robotic Systems Technology (1998) Demo III experiemental unmanned vehicle (XUV) program autonomous mobility requirements analysis. prepared for TACOM
35. Rosenblatt JK (1997) DAMN: A distributed architecture for mobile navigation. Dissertation, Carnegie Mellon
36. Rosenblatt JK (2000) Optimal selection of uncertain actions by maximizing expected utility. Auton Robots 9(1):17–25
37. Schlenoff C, Balakirsky S, Uschold M, Provine R, Smith S (2003) Using ontologies to aid navigation planning in autonomous vehicles. Knowl Eng Rev 18(3):243–255
38. Schlenoff C, Madhavan R, Barbera T (2004) A hierarchical, multi-resolutional moving object prediction approach for autonomous on-road driving. 2004 ICRA Conference, New Orleans, pp 1956–1961
39. Schlenoff C, Washington R, Barbera T, Manteuffel C (2005) A standard intelligent system ontology. Proceedings of SPIE – Unmanned Ground Vehicle Technology VII, Orlando, pp 46–56
40. Scholtz J, Antonishek B, Young J (2004) Evaluation of a human-robot interface: Development of a situational awareness methodology. National Institute of Standards and Technology, last accessed Oct 2006
41. Seares CDF (1987) Adaptive mission planning: Squeezing out greater combat capability. Air University Review, http://www.airpower.maxwell.af.mil/airchronicles/aureview/1987/seares2.html. Accessed Oct 2006
42. Touchton RA (1988) Reactor emergency action level monitor. In: Majumdar MC, Majumdar D, Sackett JI (eds) Artificial Intelligence and Other Innovative Computer Applications in the Nuclear Industry. Plenum, New York, pp 189–197
43. Touchton RA, Gunter AD, Wilson KM, Eldredge II TD, Weaver ME (1988) Reactor emergency action level monitor vol 1: REALM Technical Report. NP-5719. Electric Power Research Institute, Palo Alto
44. US Army (2003) Army Universal Task List. FM 7-15, Washington
45. W3C (2004) Web Services Architecture. World Wide Web Consortium, http://www.w3.org/TR/2004/NOTE-ws-arch-20040211/. Accessed Oct 2006
46. Weiss K, Philipps H, To TB, Kirchner A (2005) Environmental perception and situation assessment for an advanced highway assistant. In: 2005 IEEE Intelligent Vehicles Symposium Proceedings. Las Vegas, pp 472–477
47. Winston P, Horn B (1989) LISP: Third Edition. Addison-Wesley, Reading
48. Yanco HA, Drury J (2004) Where am I? Acquiring situation awareness using a remote robot platform. In: 2004 IEEE International Conference on Systems, Man and Cybernetics, The Hague, pp 2835–2840

49. Zhang W, Hill RW (2000) A template-based and pattern-driven approach to situation awareness and assessment in virtual humans. In: Fourth International Conference on Autonomous Agents, Barcelona, pp 116–123

Soliton Perturbation

Ji-Huan He
Modern Textile Institute, Donghua University,
Shanghai, China

Article Outline

Glossary

Soliton A soliton is a nonlinear pulse-like wave that can exist in some nonlinear systems. The isolated wave can propagate without dispersing its energy over a large region of space; collision of two solitons leads to unchanged forms, solitons also exhibit particlelike properties.

Soliton perturbation theory The soliton perturbation theory is used to study the solitons that are governed by the various nonlinear equations in presence of the perturbation terms.

Homotopy perturbation method The homotopy perturbation method is a useful tool to the search for solitons without the requirement of presence of small perturbations. In this method, a homotopy is constructed with a homotopy parameter, p. When $p = 0$, it becomes a nonlinear wave equation such as a KdV equation with a known soliton solution; when $p = 1$, it turns out to be the original nonlinear equation. To change p from zero to unity, one must only change from a trial soliton to the solved soliton.

Variational iteration method The variational iteration method is a new method for obtaining soliton-type solutions of various nonlinear wave equations. The method begins with a soliton-type solution with some unknown parameters which can be determined after few iterations. The iteration formulation is constructed by a general Lagrange multiplier which can be identified optimally via variational theory.

Exp-function method The exp-function method is a new method for searching for both soliton-type solutions and periodic solutions of nonlinear systems. The method assumes that the solutions can be expressed in arbitrary forms of the exp-function.

Definition of the Subject

The soliton is a kind of nonlinear wave. There are many equations of mathematical physics which have solutions of the soliton type. The first observation of this kind of wave was made in 1834 by John Scott Russell [1]. In 1895, the famous KdV equation, which possesses soliton solutions, was obtained by D. J. Korteweg and H. de Vries [2], who established a mathematical basis for the study of various solitary phenomena.

From a modern perspective, the soliton is used as a constructive element to formulate the complex dynamical behavior of wave systems throughout science: from hydrodynamics to nonlinear optics, from plasmas to shock waves, from tornados to the Great Red Spot of Jupiter, from traffic flow to the Internet, from Tsunamis to turbulence [3]. More recently, solitary waves are of key importance in the quantum fields: on extremely small scales and at very high observational resolution equivalent to a very high energy, space–time resembles a stormy ocean and particles and their interactions have soliton-type solutions [4].

Introduction

The soliton was first discovered in 1834 by John Scott Russell, who observed that a canal boat stopping suddenly gave rise to a solitary wave which traveled down the canal for several miles, without breaking up or losing strength. Russell named this phenomenon the 'soliton'.

In a highly informative as well as entertaining article [1] J.S. Russell gave an engaging historical account of the important scientific observation:

> *I was observing the motion of a boat which was rapidly drawn along a narrow channel by a pair of horses, when the boat suddenly stopped – not so the mass of water in the channel which it had put in motion; it accumulated round the prow of the vessel in a state of violent agitation, then suddenly leaving it behind, rolled forward with great velocity, assuming the form of a large solitary elevation, a rounded, smooth and well-defined heap of water, which continued its course along the channel apparently without change of form or diminution of speed. I followed it on horseback, and overtook it still rolling on at*

> *a rate of some eight or nine miles an hour, preserving its original figure some thirty feet long and a foot to a foot and a half in height. Its height gradually diminished, and after a chase of one or two miles I lost it in the windings of the channel. Such, in the month of August 1834, was my first chance interview with that singular and beautiful phenomenon which I have called the Wave of Translation.*

His ideas did not earn attention until 1965 when N.J. Zabusky and M.D. Kruskal began to use a finite difference approach to the study of KdV equation [5], and various analytical methods also led to a complete understanding of Solitons, especially the inverse scattering transform proposed by Gardner, Greene, Kruskal, and Miura [6] in 1967. The significance of Russell's discovery was then fully appreciated. It was discovered that many phenomena in physics, electronics and biology can be described by the mathematical and physical theory of the 'Soliton'.

The particle-like properties of solitons [7] also caught much attention, and were proposed as models for elementary particles [8]. More recently it has been realized that some of the quantum fields which are used to describe particles and their interactions also have solutions of the soliton type [9].

Methods for Soliton Solutions

The investigation of soliton solutions of nonlinear evolution equations plays an important role in the study of nonlinear physical phenomena. There are many analytical approaches to the search for soliton solutions, such as soliton perturbation, tanh-function method, projective approach, F-expansion method, and others [10,11,12,13,14].

Soliton Perturbation

We consider the following perturbed nonlinear evolution equation [15,16]

$$u_T + N(u) = \varepsilon R(u), \quad 0 < \varepsilon \ll 1 . \tag{1}$$

When $\varepsilon = 0$, we have the un-perturbed equation

$$u_T + N(u) = 0 , \tag{2}$$

which is assumed to have a soliton solution.

When $\varepsilon \neq 0$, but $0 < \varepsilon \ll 1$, we can use perturbation theory [15,16], and look for approximate solutions of Eq. (1), which are close to the soliton solutions of Eq. (2). Using multiple time scales (a slow time τ and a fast time t, such that $\partial_T = \partial_t + \varepsilon\partial_\tau$), we assume that the soliton solution can be expressed in the form

$$u(x, T) = u_0(\xi, \tau) + \varepsilon u_1(\xi, \tau, t) + \varepsilon^2 u_2(\xi, \tau, t) + \cdots \tag{3}$$

where $\xi = x - ct$, and τ is a slow time and t is a fast time.

Substituting Eq. (3) into Eq. (1) and then equating like-powers of ε, we can obtain a series of linear equations for u_i $(i = 0, 1, 2, 3, \ldots)$.

In most cases the nonlinear term $R(u)$ in Eq. (1) plays an import role in understanding various solitary phenomena, and the coefficient ε is not limited to a "small parameter".

Variational Approach

Recently, variational theory and homotopy technology have been successfully applied to the search for soliton solutions [17,18] without requiring the small parameter assumption. Both variational and homotopy technologies can lead to an extremely simple and elementary, but rigorous, derivation of soliton solutions.

Considering the KdV equation

$$\frac{\partial u}{\partial t} - 6u\frac{\partial u}{\partial x} + \frac{\partial^3 u}{\partial x^3} = 0 , \tag{4}$$

we seek its traveling wave solutions in the following frame

$$u(x, t) = U(\xi) \quad v(x, t) = V(\xi), \quad \xi = x - ct , \tag{5}$$

where c is angular frequency. Substituting Eq. (5) into Eq. (4) yields

$$-cu' - 6uu' + u''' = 0 , \tag{6}$$

where a prime denotes the differential with respect to ξ.

Integrating Eq. (6) yields the result

$$-cu - 3u^2 + u'' = 0 . \tag{7}$$

By the semi-inverse method [19], the following variational formulation is established

$$J = \int_0^\infty \left(\frac{1}{2}cu^2 + u^3 + \frac{1}{2}\left(\frac{\mathrm{d}u}{\mathrm{d}\xi}\right)^2 \right) \mathrm{d}\xi . \tag{8}$$

The semi-inverse method is a powerful mathematical tool to the search for variational formulae for real-life physical problems.

By the Ritz method, we search for a solitary wave solution in the form

$$u = p\,\mathrm{sech}^2(q\xi) , \tag{9}$$

where p and q are constants to be further determined.

Substituting Eq. (9) into Eq. (8) results in

$$\begin{aligned} J &= \int_0^\infty \left[\frac{1}{2}cp^2 \operatorname{sech}^4(q\xi) + p^3 \operatorname{sech}^6(q\xi) \right. \\ &\qquad \left. + \frac{1}{2}(4p^2q^2 \operatorname{sech}^4(q\xi)\tanh^2(q\xi))\right] \mathrm{d}\xi \\ &= \frac{cp^2}{2q}\int_0^\infty \operatorname{sech}^4(z)\mathrm{d}z + \frac{p^3}{q}\int_0^\infty \operatorname{sech}^6(z)\mathrm{d}z \\ &\qquad + 2p^2q\int_0^\infty \left\{\operatorname{sech}^4(z)\tanh^4(z)\right\}\mathrm{d}z \\ &= \frac{cp^2}{3q} + \frac{8p^3}{15q} + \frac{4p^2q}{15}\,. \end{aligned} \tag{10}$$

Making J stationary with respect to p and q results in

$$\frac{\partial J}{\partial p} = \frac{2cp}{3q} + \frac{24p^2}{15q} + \frac{8pq}{15} = 0\,, \tag{11}$$

$$\frac{\partial J}{\partial q} = -\frac{cp^2}{3q^2} - \frac{8p^3}{15q^2} + \frac{4p^2}{15} = 0\,, \tag{12}$$

or simplifying

$$5c + 12p + 4q^2 = 0\,, \tag{13}$$

$$-5c - 8p + 4q^2 = 0\,. \tag{14}$$

From Eqs. (13) and (14), we can easily obtain the following relations:

$$p = -\frac{1}{2}c, \quad q = \sqrt{\frac{c}{4}}\,. \tag{15}$$

So the solitary wave solution can be approximated as

$$u = -\frac{c}{2}\operatorname{sech}^2\sqrt{\frac{c}{4}}\,(x - ct - \xi_0)\,, \tag{16}$$

which is the exact solitary wave solution of KdV equation (4).

The preceding analysis has the virtue of utter simplicity. The suggested variational approach can be readily applied to the search for solitary wave solutions of other nonlinear problems, and the present example can be used as paradigms for many other applications in searching for solitary wave solutions of real-life physics problems.

Variational Iteration Method

The variational iteration method [20] is an alternative approach to soliton solutions without the requirement of establishing a variational formulation for the discussed problems [17,21,22,23,24]. As an illustrating example, we consider the $K(3,1)$ equation in the form [17]:

$$u_t + u^2u_x + u_{xxx} = 0\,. \tag{17}$$

According to the variational iteration method, its iteration formulation can be constructed as follows

$$\begin{aligned} &u_{n+1}(x,t) \\ &= u_n(x,t) - \int_0^t \left\{(u_n)_t + u_n^2(u_n)_x + (u_n)_{xxx}\right\}\mathrm{d}t\,. \end{aligned} \tag{18}$$

To search for its compacton-like solution, we assume the solution has the form

$$u_0(x,t) = \frac{a\sin^2(kx + wt)}{b + c\sin^2(kx + wt)}\,, \tag{19}$$

where a, b, k, and w are unknown constants further to be determined after few iterations [17].

Homotopy Perturbation Method

The homotopy perturbation method [25] provides a simple mathematical tool for searching for soliton solutions without any small perturbation [18,26]. Considering the following nonlinear equation

$$\frac{\partial u}{\partial t} + au\frac{\partial u}{\partial x} + b\frac{\partial^3 u}{\partial x^3} + N(u) = 0, \quad a > 0,\ b > 0\,, \tag{20}$$

we can construct a homotopy in the form

$$\begin{aligned} &(1-p)\left\{\frac{\partial u}{\partial t} + 6u\frac{\partial u}{\partial x} + \frac{\partial^3 u}{\partial x^3}\right\} \\ &\quad + p\left\{\frac{\partial u}{\partial t} + au\frac{\partial u}{\partial x} + b\frac{\partial^3 u}{\partial x^3} + N(u)\right\} = 0\,. \end{aligned} \tag{21}$$

When $p = 0$, we have

$$\frac{\partial u}{\partial t} + 6u\frac{\partial u}{\partial x} + \frac{\partial^3 u}{\partial x^3} = 0\,, \tag{22}$$

a well-known KdV equation whose soliton solution is known. When $p = 1$, Eq. (21) turns out to be the original equation. According to the homotopy perturbation method, we assume

$$u = u_0 + pu_1 + p^2u_2 + \cdots \tag{23}$$

Substituting Eq. (23) into Eq. (21), and proceeding with the same process as the traditional perturbation method does, we can easily solve u_0, u_1 and other components. The solution can be expressed finally in the form

$$u = u_0 + u_1 + u_2 + \cdots \tag{24}$$

The homotopy perturbation method always stops before the second iteration, so the solution can be expressed as

$$u = u_0 + u_1 \tag{25}$$

for most cases.

Parameter-Expansion Method

The parameter-expansion method [27,28,29,30] does not require one to construct a homotopy. To illustrate its solution procedure, we re-write Eq. (20) in the form

$$\frac{\partial u}{\partial t} + au\frac{\partial u}{\partial x} + b\frac{\partial^3 u}{\partial x^3} + 1 \cdot N(u) = 0\,. \tag{26}$$

Supposing that the parameters a, b, and 1 can be expressed in the forms

$$a = a_0 + pa_1 + p^2 a_2 + \cdots \tag{27}$$

$$b = b_0 + pb_1 + p^2 b_2 + \cdots \tag{28}$$

$$1 = pc_1 + p^2 c_2 + \cdots \tag{29}$$

where p is a bookkeeping parameter, $p = 1$. Substituting Eqs. (23), (27), (28) and (29) into Eq. (26) and proceeding the same way as the perturbation method, we can easily obtain the needed solution.

Exp-function Method

The exp-function method [31,32,33] provides us with a straightforward and concise approach to obtaining generalized solitonary solutions and periodic solutions and the solution procedure, with the help of Matlab or Mathematica, is utterly simple. Consider a general nonlinear partial differential equation of the form

$$F(u, u_x, u_y, u_z, u_t, u_{xx}, u_{yy}, u_{zz}, u_{tt}, u_{xy}, u_{xt}, u_{yt}, \ldots) = 0\,. \tag{30}$$

Using a transformation

$$\eta = ax + by + cz + dt\,, \tag{31}$$

we can re-write Eq. (30) in the form of the following nonlinear ordinary differential equation:

$$G(u, u', u'', u''', \ldots) = 0\,, \tag{32}$$

where a prime denotes a derivation with respect to η.

According to the exp-function method, the traveling wave solutions can be expressed in the form

$$u(\eta) = \frac{\sum_{n=-k}^{l} a_n \exp(n\eta)}{\sum_{m=-i}^{j} b_m \exp(m\eta)}\,, \tag{33}$$

where i, j, k, and l are positive integer which could be freely chosen, a_n and b_m are unknown constants to be determined. The solution procedure is illustrated in [32].

Future Directions

It is interesting to point out the connection of catastrophe theory to loop soliton chaos, and finally to chaotic Cantorian spacetime [34,35].

El Naschie [34,35] studied the Eguchi–Hanson gravitational instanton solution and its interpretation by 't Hooft in the context of a quantum gravitational Hilbert space, as an event and a possible solitonic "extended" particle. Transferring a certain solitonic solution of Einstein's field equations in Euclidean "real" space–time to the mathematical infinitely-dimensional Hilbert space, it is possible to observe a new non-standard process by which a definite mass can be assigned to massless particles. Thus by invoking Einstein's gravity in "solitonic" gauge theory and vice versa, an alternative explanation for how massless particles acquire mass is found, which is also in harmony with the basic structure of our standard model as it stands at present [34,35].

Bibliography

Primary Literature

1. Russell JS (1844) Report on Waves. Fourteenth Meeting of the British Association for the Advancement of Science, John Murray, London, pp 311–390
2. Korteweg DJ, De Vires G (1895) On the change of form of long waves advancing in a rectangular channel, and a new type of long stationary wave. Phil Mag Ser 539:422–443
3. Eilbeck C (2007) John Scott Russell and the solitary wave. Heriot-Watt University, Edinburgh http://www.ma.hw.ac.uk/~chris/scott_russell.html
4. El Naschie MS (2004) A review of E infinity theory and the mass spectrum of high energy particle physics. Chaos Solit Fract 19(1):209–236
5. Zabusky NJ, Kruskal MD (1965) Interaction of 'Solitons' in a collisionless plasma and the recurrence of initial states. Phys Rev Lett 15:240–243
6. Gardner CS, Greene JM, Kruskal MD, Miura RM (1967) Method for solving the KdV equation. Phys Rev Lett 19:1095–1097
7. Bode M, Liehr AW, Schenk CP et al (2002) Interaction of dissipative solitons: particle-like behavior of localized structures in a three-component reaction-diffusion system. Physica D (1–2):45–66
8. Braun HB, Kulda J, Roessli B et al (2005) Emergence of soliton chirality in a quantum antiferromagnet. Nat Phys 1(3):159–163
9. Ahufinger V, Mebrahtu A, Corbalan R et al (2007) Quantum switches and quantum memories for matter-wave lattice solitons. New J Phys 9:4
10. Ye JF, Zheng CL, Xie LS (2006) Exact solutions and localized excitations of general Nizhnik–Novikov–Veselov system in (2+1)-dimensions via a projective approach. Int J Nonlinear Sci Num Simul 7(2):203–208
11. Bogning JR, Tchakoutio-Nguetcho AS, Kofane TC (2005) Gap solitons coexisting with bright soliton in nonlinear fiber arrays.

Int J Nonlinear Sci Num Simul 6(4):371–385; Abdusalam HA (2005) On an improved complex tanh-function method. Int J Nonlinear Sci Num Simul 6(2):99–106

12. El-Sabbagh MF, Ali AT (2005) New exact solutions for (3+1)-dimensional Kadomtsev–Petviashvili equation and generalized (2+1)-dimensional Boussinesq equation. Int J Nonlinear Sci Num Simul 6(2):151–162
13. Shen JW, Xu W (2004) Bifurcations of smooth and non-smooth travelling wave solutions of the Degasperis-Procesi equation. Int J Nonlinear Sci Num Simul 5(4):397–402
14. Sheng Z (2007) Further improved F-expansion method and new exact solutions of Kadomstev–Petviashvili equation. Chaos Solit. Fract 32(4):1375–1383
15. Yu H, Yan J (2006) Direct approach of perturbation theory for kink solitons. Phys Lett A 351(1–2):97–100
16. Herman RL (2005) Exploring the connection between quasistationary and squared eigenfunction expansion techniques in soliton perturbation theory. Nonlinear Anal 63(5–7): e2473–e2482
17. He JH, Wu XH (2006) Construction of solitary solution and compacton-like solution by variational iteration method. Chaos Solit Fract 29(1):108–113
18. He JH (2005) Application of homotopy perturbation method to nonlinear wave equations. Chaos Solit Fract 26(3): 695–700
19. He JH (2004) Variational principles for some nonlinear partial differential equations with variable coefficients. Chaos Solit Fract 19(4):847–851
20. He JH (1999) Variational iteration method – a kind of non-linear analytical technique: Some examples. Int J Non-Linear Mech 34(4):699–708
21. Abulwafa EM, Abdou MA, Mahmoud AA (2007) Nonlinear fluid flows in pipe-like domain problem using variational-iteration method. Chaos Solit Fract 32(4):1384–1397
22. Inc M (2007) Exact and numerical solitons with compact support for nonlinear dispersive K(*m*,*p*) equations by the variational iteration method. Phys A 375(2):447–456
23. Soliman AA (2006) A numerical simulation and explicit solutions of KdV-Burgers' and Lax's seventh-order KdV equations. Chaos Solit Fract 29(2):294–302
24. Abdou MA, Soliman AA (2005) Variational iteration method for solving Burger's and coupled Burger's equations. J Comput Appl Math 181(2):245–251
25. He JH (2000) A coupling method of a homotopy technique and a perturbation technique for non-linear problems. Int J Non-Linear Mech 35(1):37–43
26. Ganji DD, Rafei M (2006) Solitary wave solutions for a generalized Hirota-Satsuma coupled KdV equation by homotopy perturbation method. Phys Lett A 356(2):131–137
27. Shou DH, He JH (2007) Application of Parameter-expanding Method to Strongly Nonlinear Oscillators. Int J Nonlinear Sci Numer Simul 8:113–116
28. He JH (2001) Bookkeeping parameter in perturbation methods. Int J Nonlinear Sci Numer Simul 2:257–264
29. He JH (2002) Modified Lindstedt-Poincare methods for some strongly non-linear oscillations. Part I: expansion of a constant. Int J Non-Linear Mech 37:309–314
30. Xu L (2007) He's parameter-expanding methods for strongly nonlinear oscillators. J Comput Appl Math 207(1):148–157
31. He J-H, Wu X-H (2006) Exp-function method for nonlinear wave equations. Chaos Solit Fract 30(3):700–708
32. He J-H, Abdou MA (2007) New periodic solutions for nonlinear evolution equations using Exp-function method. Chaos Solit Fract 34(5):1421–1429
33. Wu X-H, He J-H (2008) EXP-function method and its application to nonlinear equations. Chaos Solit Fract 38(3):903–910
34. El Naschie MS (2004) Gravitational instanton in Hilbert space and the mass of high energy elementary particles. Chaos Solit Fract 20(5):917–923
35. El Naschie MS (2004) How gravitational instanton could solve the mass problem of the standard model of high energy particle physics. Chaos Solit Fract 21(1):249–260

Books and Reviews

He JH (2006) Some Asymptotic Methods for Strongly Nonlinear Equations. Int J Mod Phys B 20(10):1141–1199; 20(18): 2561–2568

He JH (2006) Non-perturbative methods for strongly nonlinear problems. dissertation.de-Verlag im Internet, Berlin

Drazin PG, Johnson RS (1989) Solitons: An Introduction. Cambridge University Press, Cambridge

Solitons and Compactons

Ji-Huan He[1], Shun-dong Zhu[2]

[1] Modern Textile Institute, Donghua University, Shanghai, China

[2] Department of Science, Zhejiang Lishui University, Lishui, China

Article Outline

Glossary

Soliton A soliton is a stable pulse-like wave that can exist in some nonlinear systems. The soliton, after a collision with another soliton, eventually emerges unscathed.

Compacton A compacton is a special solitary traveling wave that, unlike a soliton, does not have exponential tails.

Generalized soliton A generalized soliton is a soliton with some free parameters. Generally a generalized soliton can be expressed by exponential functions.

Compacton-like solution A compcton-like solution is a special wave solution which can be expressed by the squares of sinusoidal or cosinoidal functions.

Definition of the Subject

Soliton and compacton are two kinds of nonlinear waves. They play an indispensable and vital role in all ramifications of science and technology, and are used as constructive elements to formulate the complex dynamical behavior of wave systems throughout science: from hydrodynamics to nonlinear optics, from plasmas to shock waves, from tornados to the Great Red Spot of Jupiter, from traffic flow to Internet, from Tsunamis to turbulence. More recently, soliton and compacton are of key importance in the quantum fields and nanotechnology especially in nanohydrodynamics.

Introduction

Solitary waves were first observed by John Scott Russell in 1895, and were studied by D. J. Korteweg and H. de Vries in 1895 . Compactons are special solitons with finite wavelength. It was Philip Rosenau and his colleagues who first found compactons in 1993. Please refer to "▶ Soliton Perturbation" for detailed information.

Solitons

A soliton is a special solitary traveling wave that after a collision with another soliton eventually emerges unscathed. Solitons are solutions of partial differential equations that model phenomena like water waves or waves along a weakly anharmonic mass-spring chain.

The Korteweg–de Vries (KdV) equation is the generic model for the study of nonlinear waves in fluid dynamics, plasma and elastic media. KdV equation is one of the most fundamental equations in nature and plays a pivotal role in nonlinear phenomena. We consider the KdV equation in the form

$$\frac{\partial u}{\partial t} + 6u\frac{\partial u}{\partial x} + \frac{\partial^3 u}{\partial x^3} = 0 \,. \tag{1}$$

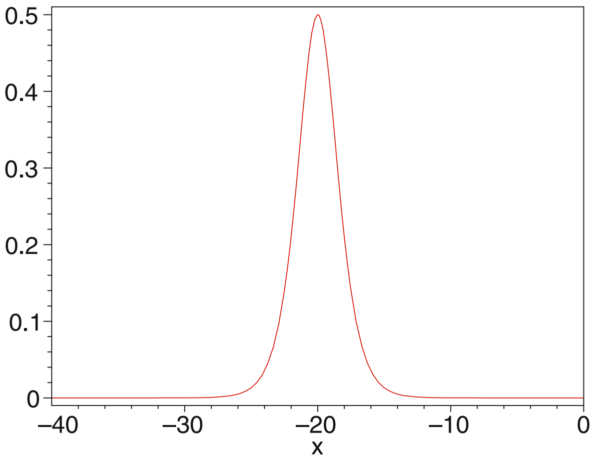
Its solitary traveling wave solution can be solved as

$$u(x,t) = \tfrac{1}{2} c \operatorname{sech}^2\left\{\tfrac{1}{2} c^{1/2}(x - ct)\right\} \,. \tag{2}$$

The bell-like solution as illustrated in Fig. 1 is called a soliton.

We re-write Eq. (2) in an equivalently form:

$$u(\xi) = p \sec h^2(q\xi) = \frac{4p}{e^{2q\xi} + e^{-2q\xi} + 2} \,. \tag{3}$$



Solitons and Compactons, Figure 1
Bell-like solitary wave

where $u(x,t) = u(\xi)$, $\xi = x - ct$, c is the wave velocity.

It is obvious that

$$\lim_{\xi\to\infty} u(\xi) = 0 \quad \text{and} \quad \lim_{\xi\to-\infty} u(\xi) = 0 \,. \tag{4}$$

The soliton has exponential tails, which are the basic character of solitary waves. The soliton obeys a superposition-like principle: solitons passing through one another emerge unmodified, see Fig. 2.

Compactons

Now consider a modified version of KdV equation in the form

$$u_t + (u^2)_x + (u^2)_{xxx} = 0 \,. \tag{5}$$

Introducing a complex variable ξ defined as $\xi = x - ct$, where c is the velocity of traveling wave, integrating once, we have

$$-cu + u^2 + (u^2)_{\xi\xi} = D \,, \tag{6}$$

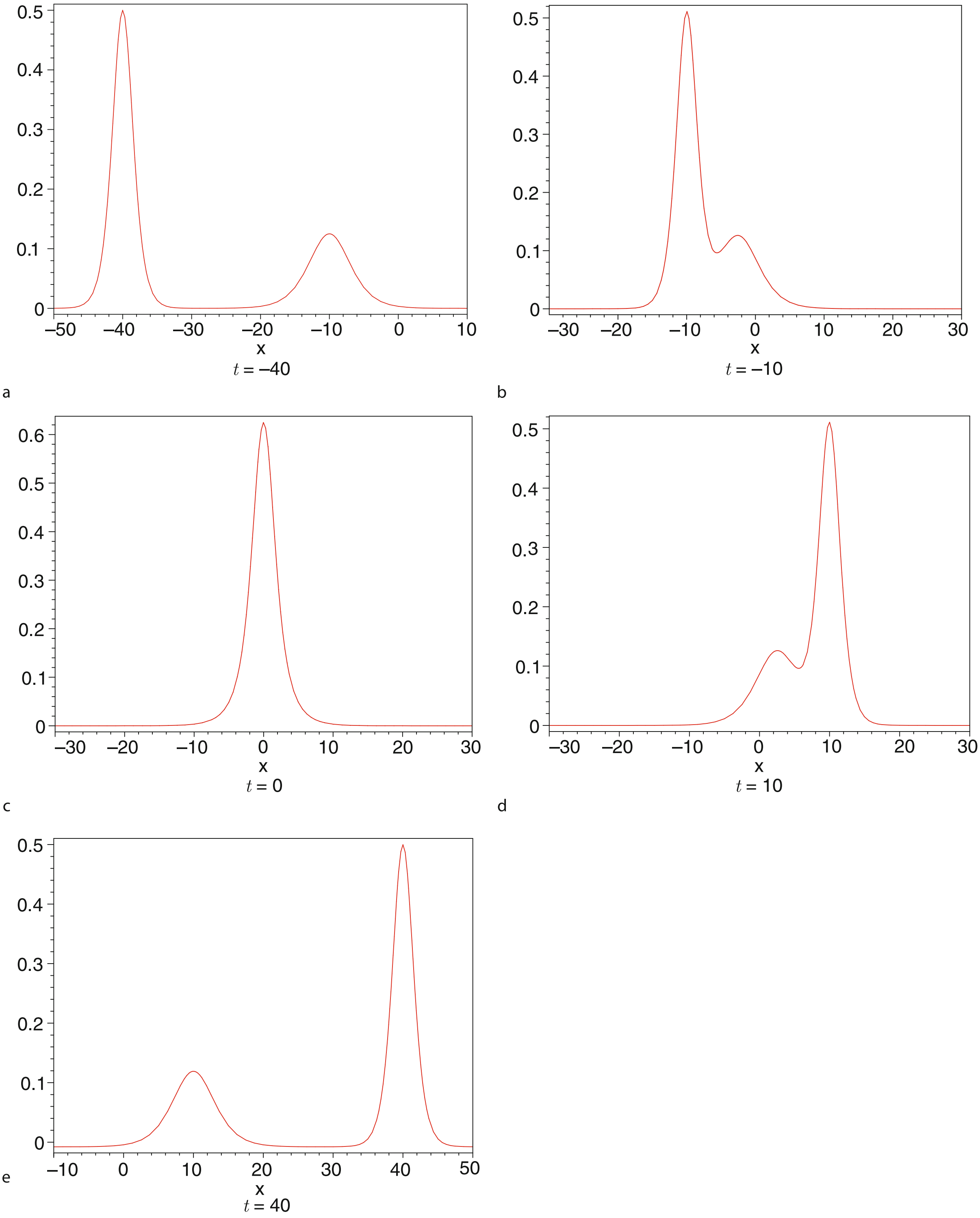
where D is an integral constant.

To avoid singular solutions, we set $D = 0$. We re-write Eq. (6) in the form

$$v_{\xi\xi} + v - cv^{1/2} = 0 \,, \tag{7}$$

where $u^2 = v$.

In case $c = 0$, we have periodic solution: $v(\xi) = A\cos\xi + B\sin\xi$. Periodic solution of nonlinear oscillators can be approximated by sinusoidal function. It helps understanding if an equation can be classified as oscillatory by direct inspection of its terms.



Solitons and Compactons, Figure 2
Collision of two solitary waves

We consider two common order differential equations whose exact solutions are important for physical understanding:

$$u'' - k^2 u = 0\,, \tag{8}$$

and

$$u'' + \omega^2 u = 0\,. \tag{9}$$

Both equations have linear terms with constant coefficients.

The crucial difference between these two very simple equations is the sign of the coefficient of u in the second term. This determines whether the solutions are exponential or oscillatory. The general solution of Eq. (8) is

$$u = Ae^{kt} + Be^{-kt}\,. \tag{10}$$

The second Eq. (9) has a positive coefficient of u, and in this case the general solution reads

$$u = A\cos\omega t + B\sin\omega t\,. \tag{11}$$

This solution describes an oscillation at the angular velocity ω.

Equation (7) behaves sometimes like an oscillator when $1 - cv^{-1/2} > 0$, i. e., $u = v^{1/2}$ has a periodic solution, we assume v can be expressed in the form

$$v = u^2 = A^2\cos^4\omega\xi\,. \tag{12}$$

Substituting Eq. (12) into Eq. (7) results in

$$\begin{aligned}12A^2\omega^2\cos^2\omega\xi - 16A^2\omega^2\cos^4\omega\xi \\ + A^2\cos^4\omega\xi - cA\cos^2\omega\xi = 0\,.\end{aligned} \tag{13}$$

We, therefore, have

$$\begin{aligned}12A^2\omega^2 - cA &= 0 \\ -16A^2\omega^2 + A^2 &= 0\,.\end{aligned} \tag{14}$$

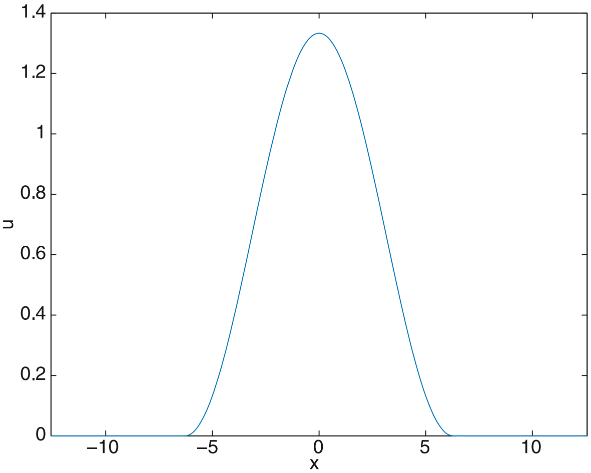
Solving the above system, Eq. (14), yields

$$\omega = \frac{1}{4}\,, \quad A = \frac{4}{3}c\,. \tag{15}$$

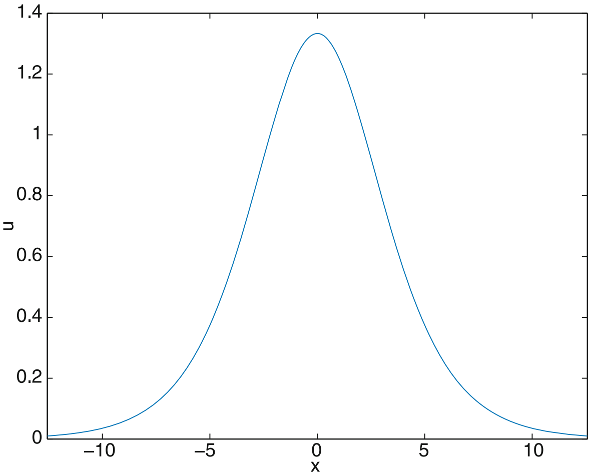
We obtain the solution in the form

$$u = v^{1/2} = \frac{4c}{3}\cos^2\left[\frac{1}{4}(x-ct)\right]\,. \tag{16}$$

By a careful inspection, v can tend to a very small value or even zero, as a result, $1 - cv^{-1/2}$ tends to negative infinite, and Eq. (7) behaves like Eq. (8) with $k \to \infty$, the exponential tails vanish completely at the edge of the bell-shape (see Fig. 3):



Solitons and Compactons, Figure 3
Compaton wave without tails



Solitons and Compactons, Figure 4
Solitary wave with two tails

$$u = \begin{cases}\frac{4c}{3}\cos^2\left[\frac{1}{4}(x-ct)\right]\,, & |x-ct| \le 2\pi \\ 0\,, & \text{otherwise.}\end{cases} \tag{17}$$

This is a compact wave. Unlike solitons (Fig. 4), compacton does not have exponential tails (Fig. 3).

Generalized Solitons and Compacton-like Solutions

Solitary solutions have tails, which can be best expressed by exponential functions. We can assume that a solitary solution can be expressed in the following general form

$$u(\eta) = \frac{\sum\limits_{n=-c}^{d} a_n \exp(n\eta)}{\sum\limits_{m=-p}^{q} b_m \exp(m\eta)}\,, \tag{18}$$

where c, d, p, and q are positive integers which are unknown to be further determined, a_n and b_m are unknown constants. The unknown constants can be easily determined using Matlab, the method is called the Exp-function method.

We consider the modified KdV equation in the form:

$$u_t + u^2 u_x + u_{xxx} = 0\,. \tag{19}$$

Using a transformation: $u(x, t) = u(\xi), \xi = kx + \omega t$, we have

$$\omega u' + ku^2 u' + k^3 u''' = 0\,, \tag{20}$$

where prime denotes the differential with respect to ξ.

We suppose that the solution of Eq. (20) can be expressed as

$$u(\xi) = \frac{a_c \exp(c\xi) + \cdots + a_{-d} \exp(-d\xi)}{b_p \exp(p\xi) + \cdots + b_{-q} \exp(-q\xi)}\,. \tag{21}$$

To determine values of c, d, p and q, we balance the linear term of highest order in Eq. (20) with the highest order nonlinear term. According to the homogeneous balance principle, we obtain the result $c = p$ and $d = q$. For simplicity, we set $c = p = 1$ and $d = q = 1$, so Eq. (21) reduces to

$$u(\xi) = \frac{a_1 \exp(\xi) + a_0 + a_{-1} \exp(-\xi)}{\exp(p\xi) + b_0 + b_{-1} \exp(-\xi)}\,. \tag{22}$$

Substituting Eq. (22) into Eq. (20), and by the help of Matlab, clearing the denominator and setting the coefficients of power terms like $\exp(j\xi), j = 1, 2, \cdots$ to zero yield a system of algebraic equations, solving the obtained system, we obtain the following exact solutions:

$$\begin{cases} a_0 = a_1 b_0 + \dfrac{3k^2 b_0}{a_1}\,, & a_{-1} = \dfrac{b_0^2\left(3k^2 + 2a_1^2\right)}{8a_1}\,, \\ b_{-1} = \dfrac{b_0^2\left(3k^2 + 2a_1^2\right)}{8a_1^2}\,, & \omega = -ka_1^2 - k^3\,, \end{cases} \tag{23}$$

where a_1 and b_0 are free parameters, which depends upon the initial conditions and/or boundary conditions. The property that stability may depend on initial/boundary conditions is characteristic only for nonlinear systems.

The relationship between wave speed and frequency is

$$\omega = -ka_1^2 - k^3\,. \tag{24}$$

Note that the value of a_1 is determined from the initial/boundary conditions, so frequency or wave speed may not independent of initial/boundary conditions.

Then, the closed form solution of Eq. (19) reads

$$\begin{aligned} &u(x, t) \\ &= \frac{\left(a_1 \exp[kx - (ka_1^2 + k^3)t] + a_1 b_0 + \frac{3k^2 b_0}{a_1} + \frac{b_0^2(3k^2+2a_1^2)}{8a_1} \exp[-kx + (ka_1^2 + k^3)t]\right)}{\left(\exp[kx - (ka_1^2 + k^3)t] + b_0 + \frac{b_0^2(3k^2+2a_1^2)}{8a_1^2} \exp[-kx + (ka_1^2 + k^3)t]\right)} \\ &= a_1 + \frac{\frac{3k^2 b_0}{8}}{\left(\exp[kx - (ka_1^2 + k^3)t] + b_0 + \frac{b_0^2(3k^2+2a_1^2)}{8a_1^2} \exp[-kx + (ka_1^2 + k^3)t]\right)}\,. \end{aligned} \tag{25}$$

Generally a_1, b_0 and k are real numbers, and the obtained solution, Eq. (25), is a generalized soliton solution.

If we choose $k = 1, a_1 = 1, b_0 = \sqrt{8/5}$, Eq. (25) becomes

$$u(x, t) = 1 + \frac{3\sqrt{1/40}}{\exp[x - 2t] + \sqrt{8/5} + \exp[-x + 2t]}\,. \tag{26}$$

The bell-like solution is illustrated in Fig. 5.

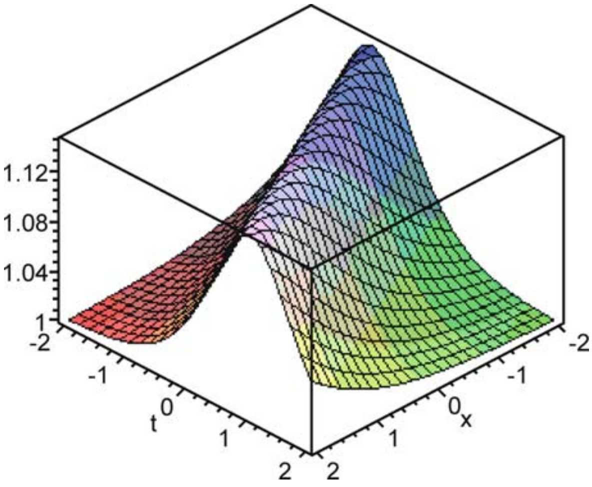
In case k is an imaginary number, the obtained solitary solution can be converted into periodic solution or compact-like solution. We write $k = iK$, Eq. (25) becomes

$$\begin{aligned} &u(x, t) = a_1 \\ &+ \frac{-\frac{3K^2 b_0}{8}}{\left((1 + p) \exp[Kx - (Ka_1^2 - K^3)t] + b_0 + i(1 - p)\frac{b_0^2(3k^2+2a_1^2)}{8a_1^2} \exp[-Kx + (Ka_1^2 - K^3)t]\right)}\,, \end{aligned} \tag{27}$$

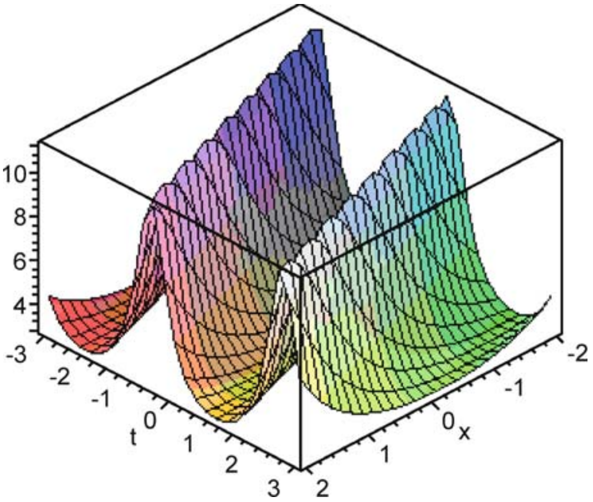
where $p = \frac{b_0^2(-3K^2+2a_1^2)}{8a_1^2}$.

If we search for a periodic solution or compact-like solution, the imaginary part in the denominator of Eq. (27) must be zero, that requires that

$$1 - p = 1 - \frac{b_0^2(-3K^2 + 2a_1^2)}{8a_1^2} = 0\,. \tag{28}$$



Solitons and Compactons, Figure 5
Propagation of a solution with respect to time



Solitons and Compactons, Figure 6
Periodic solution

Solving b_0 from Eq. (28), we obtain

$$b_0 = \pm\sqrt{\frac{8}{-3K^2 + 2a_1^2}}\,. \tag{29}$$

Substituting Eq. (29) into Eq. (27) results in a periodic solution, which reads

$$u(x,t) = a_1 + \frac{\pm 3K^2\sqrt{\frac{2}{-3K^2+2a_1^2}}}{\cos[Kx - (Ka_1^2 - K^3)t] \pm \sqrt{\frac{2}{-3K^2+2a_1^2}}} \tag{30}$$

or a generalized compact-like solution:

$$u(x,t) = \begin{cases} a_1 + \dfrac{\pm 3K^2\sqrt{\frac{2}{-3K^2+2a_1^2}}}{\left(\cos[Kx-(Ka_1^2-K^3)t] \pm \sqrt{\frac{2}{-3K^2+2a_1^2}}\right)}\,, \\ a_1 + 3K^2\,, \quad \text{otherwise} \end{cases}$$
$$\left|Kx - \left(Ka_1^2 - K^3\right)t\right| \le \frac{\pi}{2} \tag{31}$$

where a_1 and K are free parameters, and it requires that $2a_1^2 > 3K^2$. If we choose $k = 1, a_1 = 1, b_0 = \sqrt{8/5}$, Eq. (30) becomes

$$u(x,t) = 1 + \frac{3\sqrt{2}}{\cos[x - 3t] + \sqrt{2}}\,. \tag{32}$$

The periodic solution is illustrated in Fig. 6.

Now we give an heuristical explanation of why Eq. (19) behaves sometimes periodically and sometimes compaton-like.

We re-written Eq. (20) in form

$$u'' + \frac{\omega}{k^3}u + \frac{1}{3k^2}u^3 = 0. \tag{33}$$

It is a well-known Duffing equation with a periodic solution for all $\omega > 0$ and $k > 0$.

Actually in our study ω can be negative, we re-write Eq. (33) in the form

$$u'' - \frac{\omega}{k^3}u + \frac{1}{3k^2}u^3 = 0\,, \quad \omega > 0\,. \tag{34}$$

This equation, however, has not always a periodic solution. We use the parameter-expansion method to find its period and the condition to be an oscillator. In order to carry out a straightforward expansion like that in the classical perturbation method, we need to introduce a parameter, λ, because none appear explicitly in this equation. To this end, we seek an expansion in the form

$$u = u_0 + \lambda u_1 + \lambda^2 u_2 + \lambda^3 u_3 + \cdots\,. \tag{35}$$

The parameter λ is used as a bookkeeping device and is set equal to unity.

The coefficients of the linear term and nonlinear term can be, respectively, expanded in a similar way:

$$-\frac{\omega}{k^3} = \Omega^2 + m_1\lambda + m_2\lambda^2 + \ldots \tag{36}$$

$$\frac{1}{3k^2} = n_1\lambda + n_2\lambda^2 + \ldots\,, \tag{37}$$

where m_i and n_i are unknown constants to be further determined.

Interpretation of why such expansions work well is given by [1].

Substituting Eqs.(35)–(37) to (34), we have

$$\begin{aligned}&\left(u_0+\lambda u_1+\lambda^2 u_2+\dots\right)''\\+&\left(\Omega^2+m_1\lambda+m_2\lambda^2+\dots\right)\left(u_0+\lambda u_1+\lambda^2 u_2+\dots\right)\\+&\left(n_1\lambda+n_2\lambda^2+\dots\right)\cdot\left(u_0+\lambda u_1+\lambda^2 u_2+\dots\right)^3=0\end{aligned} \tag{38}$$

and equating coefficients of like powers of λ, we obtain

Coefficient of λ^0

$$u_0''+\Omega^2 u_0=0\,. \tag{39}$$

Coefficient of λ^1

$$u_1''+\Omega^2 u_1+m_1 u_0+n_0 u_0^3=0\,. \tag{40}$$

The solution of Eq. (39) is

$$u_0=A\cos\Omega t\,. \tag{41}$$

Substituting u_0 into (40) gives

$$\begin{aligned}u_1''+\Omega^2 u_1+A(m_1+\tfrac{3}{4}n_0A^2)\cos\Omega t&\\+\tfrac{1}{4}n_0A^3\cos 3\Omega t=0\,.&\end{aligned} \tag{42}$$

No secular term in u_1 requires that

$$m_1+\frac{3}{4}n_0A^2=0\quad\text{or}\quad A=0\,. \tag{43}$$

If the first-order approximate solution is searched for, then we have

$$-\frac{\omega}{k^3}=\Omega^2+m_1 \tag{44}$$

$$\frac{1}{3k^2}=n_1\,. \tag{45}$$

We finally obtain the following relationship

$$\Omega^2=-\frac{\omega}{k^3}+\frac{1}{4k^2}A^2\,. \tag{46}$$

To behave like an oscillator requires that

$$-\frac{\omega}{k^3}+\frac{1}{4k^2}A^2>0 \tag{47}$$

or

$$\frac{\omega}{k}<\frac{1}{4}A^2\,. \tag{48}$$

The amplitude A may strongly depend upon initial/boundary conditions which may determine the wave type of a nonlinear equation.

Now we approximate Eq. (34) in the form

$$u''+\frac{1}{k^2}\left(-\frac{\omega}{k}+\frac{A^2\cos^2\Omega t}{3}\right)u=0\,. \tag{49}$$

In case $|\Omega t|\to\pi/2$, the above equation behaves exponentially, resulting in a compact-like wave as discussed above.

Future Directions

It is interesting to identify the conditions for a nonlinear equation to have solitary, or periodic, or compacton-like solutions. In most open literature, many papers on soliton and compacton are focused themselves on a special solution with either a soliton or a compacton without considering the initial/boundary conditions, which might be vital important for its actual wave type.

Solitons and compactons for difference-differential equations (e. g. Lotka–Volterra-like problems) have been caught much attention due to the fact that discrete space-time may be the most radical and logical viewpoint of reality (refer to E-infinity theory detailed concept). For small scales, e. g., nano scales, the continuum assumption becomes invalid, and difference equations have to be used for space variables.

Fractional differential model is another compromise between the discrete and the continuum, and can best describe solitons and compactons.

Many interesting phenomena arise in nanohydrodynamics recently, such as remarkably excellent thermal and electric conductivity, and extremely extraordinary fast flow in nanotubes. Consider a single compacton wave along a nanotube, and its wavelength is as same as the diameter of the nanotubes, under such a case, almost no energy is lost during the transportation, resulting in extremely extraordinary fast flow in the nanotubes.

The physical understanding of the transformation $k=iK$ is also worth further studying.

Cross References

▶ Soliton Perturbation

Bibliography

Primary Literature

1. He JH (2006) New interpretation of homotopy perturbation method. Int J Mod Phys 20(18):2561–2568

Some Famous Papers on Solitons and Compactons

2. Burger S, Bongs K, Dettmer S et al (1999) Dark solitons in Bose–Einstein condensates. Phys Rev Lett 83(25):5198–5201
3. Denschlag J, Simsarian JE, Feder DL et al (2000) Generating solitons by phase engineering of a Bose–Einstein condensate. Science 287(5450):97–101
4. Diakonov D, Petrov V, Polyakov M (1997) Exotic anti-decuplet of baryons: prediction from chiral solitons. Z Phys A 359(3):305–314
5. Duff MJ, Khuri RR, Lu JX (1995) Sting Solitons. Phys Rep 259(4–5):213–326
6. Eisenberg HS, Silberberg Y, Morandotti R et al (1998) Discrete spatial optical solitons in waveguide arrays. Phys Rev Lett 81(16):3383–3386
7. Fleischer JW, Segev M, Efremidis NK, et al (2003) Observation of two-dimensional discrete solitons in optically induced nonlinear photonic lattices. Nature 422(6928):147–150
8. He JH (2005) Application of homotopy perturbation method to nonlinear wave equations. Chaos Soliton Fract 26(3):695–700
9. He JH, Wu XH (2006) Construction of solitary solution and compacton-like solution by variational iteration method. Chaos Soliton Fract 29(1):108–113
10. Khaykovich L, Schreck F, Ferrari G et al (2002) Formation of a matter-wave bright soliton. Science 296(5571):1290–1293
11. Rosenau P (1997) On nonanalytic solitary waves formed by a nonlinear dispersion. Phys Lett A 230(5–6):305–318
12. Rosenau P (2000) Compact and noncompact dispersive patterns. Phys Lett A 275(3):193–203
13. Strecker KE, Partridge GB, Truscott AG et al (2002) Formation and propagation of matter-wave soliton trains. Nature 417(6885):150–153
14. Torruellas WE, Wang Z, Hagan DJ et al (1995) Observation of 2-dimensional spatial solitary waves in a quadratic medium. Phys Rev Lett 74(25):5036–5039

Review Article

15. He JH (2006) Some asymptotic methods for strongly nonlinear equations. Int J Mod Phys B 20(10):1141–1199 and 20(18):2561–2568

Exp-function Method

16. He JH, Wu XH (2006) Construction of solitary solution and compacton-like solution by variational iteration method. Chaos Soliton Fract 29(1):108–113
17. He JH, Wu XH (2006) Exp-function method for nonlinear wave equations. Chaos Soliton Fract 30(3):700–708
18. Zhu SD (2007) Exp-function method for the Hybrid–Lattice system. Int J Nonlinear Sci 8(3):461–464
19. Zhu SD (2007) Exp-function method for the discrete mKdV lattice. Int J Nonlinear Sci 8(3):465–468

Parameter-Expansion Method

20. He JH (2001) Bookkeeping parameter in perturbation methods. Int J Nonlinear Sci Numer Simul 2(3):257–264
21. He JH (2002) Modified Lindstedt–Poincare methods for some strongly non-linear oscillations Part I: Expansion of a constant. Int J Nonlinear Mech 37(2):309–314
22. He JH (2006) Non-perturbative methods for strongly nonlinear problems. dissertation.de-Verlag im Internet GmbH, Berlin
23. Shou DH, He JH (2007) Application of parameter-expanding method to strongly nonlinear oscillators. Int J Nonlinear Sci Numer Simul 8(1):121–124
24. Xu L (2007) Application of He's parameter-expansion method to an oscillation of a mass attached to a stretched elastic wire. Phys Lett A 368(3–4):259–262
25. Xu L (2007) Determination of limit cycle by He's parameter-expanding method for strongly nonlinear oscillators. J Sound Vib 302(1–2):178–184

Nanohydrodynamics and Nano-effect

26. He JH, Wan Y-Q, Xu L (2007) Nano-effects, quantum-like properties in electrospun nanofibers. Chaos Solitons Fract 33(1), 26–37
27. Majumder M, Chopra N, Andrews R et al (2005) Nanoscale hydrodynamics – Enhanced flow in carbon nanotubes. Nature 438(7064):44–44

E-Infinity Theory

28. El Naschie MS (2007) A review of applications and results of E-infinity theory. Int J Nonlinear Sci Numer Simul 8(1):11–20
29. El Naschie MS (2007) Deterministic quantum mechanics versus classical mechanical indeterminism. Int J Nonlinear Sci Numer Simul 8(1):5–10

Fractional-Order Differential Equations

30. Draganescu GE (2006) Application of a variational iteration method to linear and nonlinear viscoelastic models with fractional derivatives. J Math Phys 47(8):082902
31. He JH (1998) Approximate analytical solution for seepage flow with fractional derivatives in porous media. Comput Methods Appl Mech Eng 167(1–2):57–68
32. Momani S, Odibat Z (2007) Homotopy perturbation method for nonlinear partial differential equations of fractional order. Phys Lett A 365(5–6):345–350
33. Odibat ZM, Momani S (2006) Application of variational iteration method to Nonlinear differential equations of fractional order. Int J Nonlinear Sci Numer Simul 7(1):27–34
34. Wang Q (2007) Homotopy perturbation method for fractional KdV equation. Appl Math Comput 190(2):1795–1802

Solitons: Historical and Physical Introduction

FRANÇOIS MARIN
Laboratoire Ondes et Milieux Complexes, Fre CNRS
3102, Le Havre Cedex, France

Article Outline

Glossary

Breaking waves As waves increase in height through the shoaling process, the crest of the wave tends to speed up relative to the rest of the wave. Waves break when the speed of the crest exceeds the speed of the advance of the wave as a whole.

Crystal lattice A geometric arrangement of the points in space at which the atoms, molecules, or ions of a crystal occur.

Deep water Water sufficiently deep that surface waves are little affected by the ocean bottom. Water deeper than one-half the surface wave length is considered deepwater.

Fluxon Quantum of magnetic flux.

Freak waves Single waves which result from a local focusing of wave energy. They are of considerable danger to mariners because of their unexpected nature.

Geostrophic adjustment The process by which an unbalanced atmospheric flow field is modified to geostrophic equilibrium, generally by a mutual adjustment of the atmospheric wind and pressure fields depending on the initial horizontal scale of the disturbance.

Geostrophic equilibrium A state of motion of an inviscid fluid in which the horizontal Coriolis force exactly balances the horizontal pressure force at all points of the field.

Hydraulic jump A sudden turbulent rise in water level, such as often occurs at the foot of a spillway when the velocity of rapidly flowing water is instantaneously slowed.

Katabatic wind Most widely used in mountain meteorology to denote a downslope flow driven by cooling at the slope surface during periods of light larger-scale winds.

Lightning Lightning is a transient, high-current electric discharge.

Plasma Hot, ionized gas.

Shallow water Water depths less than or equal to one half of the wavelength of a wave.

Solitary wave Localized wave that propagates along one space direction only, with undeformed shape.

Soliton Large-amplitude pulse of permanent form whose shape and speed are not altered by collision with other solitary waves, the exact solution of a nonlinear equation.

Spillway A feature in a dam allowing excess water to pass without overtopping the dam.

Thermocline A layer in which the temperature decreases significantly (relative to the layers above and below) with depth.

Synoptic scale Used with respect to weather systems ranging in size from several hundred kilometers to several thousand kilometers.

Thunder The sound emitted by rapidly expanding gases along the channel of a lightning discharge.

Thunderstorm In general, a local storm, invariably produced by a cumulonimbus cloud and always accompanied by lightning and thunder, usually with strong gusts of wind, heavy rain, and sometimes with hail.

Tidal bore Tidal wave that propagates up a relatively shallow and sloping estuary or river, in a solitary wave form. The leading edge presents an abrupt rise in level, sometimes with continuous breaking and often immediately followed by several large undulations. The tidal bore is usually associated with high tidal range and a sharp narrowing and shoaling at the entrance. Also called pororoca (Brazilian) and mascaret (French).

Troposphere The portion of the atmosphere from the earth's surface to the tropopause, that is the lowest 10–20 km of the atmosphere.

Tsunami Long period ocean wave generated by an earthquake or a volcanic explosion.

Definition of the Subject

The interest in nonlinear physics has grown significantly over the last fifty years. Although numerous nonlinear processes had been previously identified the mathematic tools of nonlinear physics had not yet been developed. The available tools were linear, and nonlinearities were avoided or treated as perturbations of linear theories. The solitary water wave, experimentally discovered in 1834 by John Scott Russell, led to numerous discussions. This hump-shape localized wave that propagates along one space-direction with undeformed shape has spectacular stability properties. John Scott Russell carried out many experiments to obtain the properties of this wave. The theories which were based on linear approaches concluded that this kind of wave could not exist. The controversy was resolved by J. Boussinesq [5] and by Lord Rayleigh [64] who showed that if dissipation is neglected, the increase in local wave velocity associated with finite amplitude is balanced by the decrease associated with dispersion, leading to a wave of permanent form. A model equation describing the unidirectional propagation of long waves in water of relatively shallow depth with a localized solution representing a single hump as discovered by Russell was obtained by Korteweg and de Vries [42]. This equation has become very famous and is now known as the Korteweg–de Vries equation or KdV equation. These results were not considered very important when they were obtained. A remarkable numerical discovery was made by E. Fermi, J. Pasta and S. Ulam [18] as they studied the flow of incoherent energy in a solid modeled by a one-dimensional lattice of equal masses connected by nonlinear springs. The initial injected energy was not shared among all the degrees of freedom of the lattice, but returned almost entirely to the original excited mode. The explanation was given only ten years later by Zabusky and Kruskal [82] from numerical solutions of the KdV equation used to model in the continuum approximation a nonlinear atomic lattice with periodic boundary conditions. This led to the concept of solitons and ultimately to the development of integrable systems. The term soliton was chosen as it is a localized wave which propagates preserving its shape and velocity as a particle would propagate. A soliton is then a large-amplitude pulse of permanent form whose shape and speed are not altered by collision with other solitary waves, and it is the exact solution of a nonlinear equation. The mathematical features of solitons are often well developed in soliton literature since they are at the origin of very nice theoretical developments such as the powerful inverse scattering transform which solves a complex nonlinear equation through a series of linear steps. Nevertheless, the physics of solitons is very rich and it is a very actual research topic in numerous fields, for example in hydrodynamics, in optics, in electricity, in solid physics, in chemistry and in biology. Nonlinearity and dispersion are very common in macroscopic physics as well as in microscopic physics. Nonlinearity tends to localize the signals while dispersion spreads them. These two opposite effects sometimes compensate and the regime of solitons takes place. Real physical systems are only approximately described by the equations of the theory of solitons, but a remarkable feature of solitons is their very high stability relative to perturbations. The soliton concept is now firmly established. They are widely accepted as a structural basis for viewing and understanding the dynamic behavior of complex nonlinear systems.

Introduction

The first experimental observation of a solitary wave was made in August 1834 by a Scottish engineer named John Scott Russell (1808–1882). He was mounted on a horse along the Union Canal linking Edinburgh with Glasgow when he saw a rounded smooth well-defined heap detach itself from the prow of a boat brought to rest and continue its course without change of shape or diminution of speed. Scott–Russell reported his observation in the British Association Report [71] as follows:

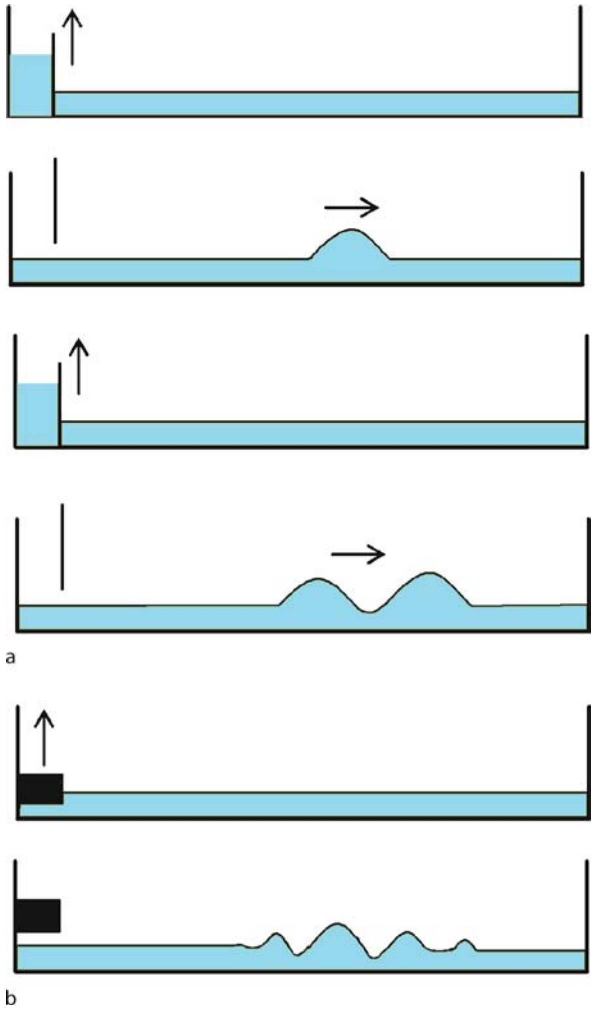
"I was observing the motion of a boat which was rapidly drawn along a narrow channel by a pair of horses, when the boat suddenly stopped—not so the mass of water in the channel which it had put in motion; it accumulated round the prow of the vessel in a state of violent agitation, then suddenly leaving it behind, rolled forward with great velocity, assuming the form of a large solitary elevation, a rounded, smooth and well-defined heap of water, which continued its course along the channel apparently without change of form or diminution of speed. I followed it on horseback, and overtook it still rolling on at a rate of some eight or nine miles an hour, preserving its original figure some thirty feet long and a foot to a foot and a half in height. Its height gradually diminished, and after a chase of one or two miles I lost it in the windings of the channel. Such, in the month of August 1834, was my first chance interview with that singular and beautiful phenomenon which I have called the Wave of Translation".

Russell wrote that this August day in 1834 was the happiest in his life. In July 1995, an international gathering of scientists witnessed a re-creation of the solitary wave observed by Russell on an aqueduct of the Union Canal. This aqueduct was named after this occasion *Scott Russell aqueduct*; the scientists were attending a conference on nonlinear waves in physics and biology at Heriot-Watt University (Edinburgh), near the canal.

Throughout his life Russell remained convinced that "his" wave that he initially called the "Wave of Translation" and later the "Great Solitary Wave" was of fundamental importance. He therefore carried out extensive experiments in wave-tanks and established the velocity formula for solitary waves:

$$V_S = \sqrt{g(d + A_S)} \tag{1}$$

where g is the acceleration due to gravity, d the undisturbed water depth and A_S the amplitude of the solitary wave. The "quasi-cycloidal" form of the wave was found to be independent of the way it was generated. Russell [69] described the induced movement of fluid particles: "By the transit of the wave the particles of the fluid are raised from their places, transferred forwards in the direction of the motion of the wave, and permanently deposited at rest in a new place at a considerable distance from their original position". He also deduced from his numerous experiments that two solitary waves can cross



Solitons: Historical and Physical Introduction, Figure 1
Schematic view of Scott Russell's experiments

each other "without change of any kind". Figure 1 shows a schematic view of his experiments in tanks, adapted from Remoissenet [65]. An initial elevation of water may induce one or two solitary waves depending on the relation between its height and length (Fig. 1a); the case where an initial depression does not evolve into solitary waves but leads to an oscillatory wave train of gradually increasing length and decreasing amplitude is shown in Fig. 1b.

In Sect. "Historical Discovery of Solitons", the historical discovery of solitons is presented. The physical concept of solitons and the associated applications are described in Sect. "Physical Properties of Solitons and Associated Applications". In Sect. "Mathematical Methods Suitable for the Study of Solitons", the mathematical methods suitable for the study of solitons are considered. Finally, Sect. "Future Directions" is devoted to future directions.

Historical Discovery of Solitons

After Russell had carried out his observation of the solitary wave in August 1834, he was put in charge by the Union Canal Company of determining the efficiency of canals for the transport of steam-driven barges. He worked on the resistance of boats towed through water of finite depth. A resistance proportional to the square of the velocity V_b of the boat was predicted by the theory at that time. Convinced of the imperfection of this theory, Russell performed numerous experiments in the summers of 1834 and 1835 [12]. Several boats were towed in canals of various depths at velocities ranging from 3 to 15 miles per hour. The resistance was measured by a dynamometer. Resistance curves such as the one depicted in Fig. 2 were obtained, with a resistance local maximum R_m for a boat velocity V_{cr} and a growing deviation from the theoretical curve for increasing values of boat velocity. The critical velocity V_{cr} was found to depend on the water depth. Russell noticed that this critical velocity was the same as the velocity of the solitary wave for the given water depth d, and gave a qualitative explanation for the diminution of flow resistance for increasing values of the boat velocity just above the critical velocity V_{cr}. For this description, he considered the variation of the shape of the water surface around the moving boat with the velocity. For velocities smaller than the critical value, the water level is raised around the prow where a wave of displacement is generated, leading to an inclination of the boat. The resulting effect is an augmentation of the section of immersion and of the corresponding flow resistance. When the velocity of the boat reaches the critical value, the wave of displacement has the velocity of a solitary wave and the vessel travels on this solitary wave. For velocities greater than the critical value, the boat stays on the wave summit leading to a smaller flow resistance. Russell [70] tried a quantitative explanation of this process, but this paper contained several errors. However, Russell was a very fine observer; he was far ahead of his time in considering the fundamental importance of solitary waves. He mentioned [70] that "a large or high wave had a greater velocity than a small one. When a small wave preceded a large one, the latter invariably overtook the other, and when the large wave was before the less, their mutual distance invariably became greater". He showed that the shape of solitary waves depends on their height, the width of the wave decreasing for increasing values of height. It is also fascinating that Russell had observed the collision properties of solitary waves [8]. Nevertheless, the observations of Russell induced numerous critical comments. Airy [2] wrote in his treatise "Tides and Waves": "We are not disposed to recognize this wave (discovered by Scott

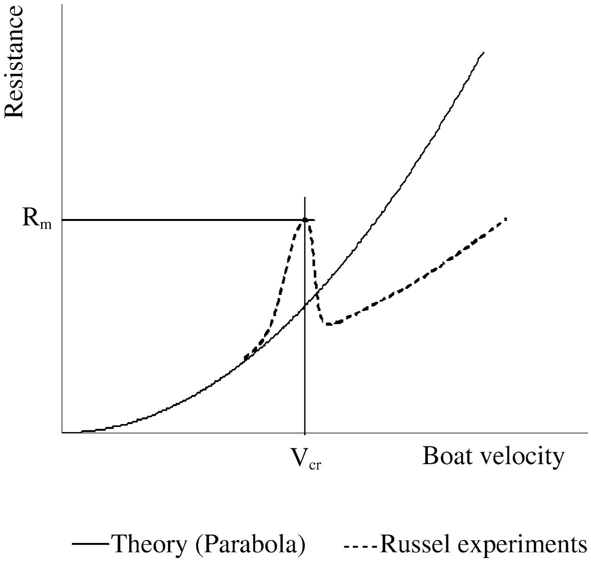
Russell) as deserving the epithets "great" or "primary", and we conceive that ever since it was known that the theory of shallow waves of great length was contained in the equation $\partial^2 X/\partial t^2 = gd\partial^2 X/\partial x^2$ the theory of the solitary wave has been perfectly well known". In the equation quoted by Airy, t is the time and x is the horizontal coordinate of a fluid particle. The wave velocity in shallow water at the first order of approximation, when the ratio H/d between the wave height H and the water depth may be neglected is $V_0 = \sqrt{gd}$, as found by Lagrange in 1786; this velocity differs from the formulation proposed by Russell for solitary waves (Eq. (1)). Airy mentioned further: "Some experiments were made by Mr. Russell on what he calls a negative wave. But (we know not why) he appears not to have been satisfied with these experiments and had omitted them in his abstract. All of the theorems of our IVth section, without exception, apply to these as well as to positive waves, the sign of the coefficient only being changed". Also, according to Airy [2] and Lamb [46], long waves could not propagate in a canal of rectangular section without changing their shape. The controversy arose because the dispersion is neglected by the nonlinear shallow water theory. Stokes [73] considered Russell results and tried to obtain them analytically; however, his linear or weakly nonlinear approach did not permit him to retrieve the results of Russell, which made him doubtful about them. In order to distinguish his "great wave" from other waves, Russell [71] introduced four orders of waves:

1. Waves of translation. These waves involve mass transfer. The solitary waves observed by Russell are included in this order.
2. Oscillatory waves. These are the waves which can be the most often observed; they do not involve mass transfer.
3. Capillary waves. The surface tension effects are important for those waves.
4. Corpuscular waves. Those waves are rapid successions of solitary waves.

Russell was most concerned with the first order, but he also carried out experiments with waves of the second and third order. The fourth order (corpuscular waves) suggested by Russell had been seriously criticized, and the manuscripts he submitted on this order were never published due to ignorance of mechanics principles.

The great difficulties encountered by Russell to prove the importance of his findings strongly disappointed him. Tired of discussions, he stopped his work on solitary waves and started to build large steam ships with great success. A detailed biography of Russell may be found in [16].

A French mathematician, Joseph Boussinesq, knew of the existence of Russell's observations, and also of detailed experiments performed by Bazin [3] in the long branch of the canal de Bourgogne close to Dijon (France), which confirmed Russell's observations. Boussinesq tried to obtain a solution of Euler's equations compatible with Russell's and Bazin's results. He developed the horizontal and vertical velocity components of fluid velocity in a rectangular channel in power of the distance from the bed. Including nonlinear terms which had been neglected by Lagrange, Boussinesq [5] obtained a solution with the properties of the solitary wave observed by Russell. In particular, the shape of the wave was found to be given by:



Solitons: Historical and Physical Introduction, Figure 2
Resistance as a function of towing velocity according to Russell

$$\eta = A_S \operatorname{sech}^2 \left[\sqrt{\frac{3A_S}{4d^3}} (x - V_S t) \right] \tag{2}$$

where η is the free surface displacement and $\operatorname{sech} x = 1/\cosh x$; the velocity of the wave V_S is given by Eq. (1). The article of Boussinesq [5] had been almost ignored in England, and five years after its publication Lord Rayleigh [64] independently obtained the solitary wave profile. Following Rayleigh's work and including the effects of surface tension, the Amsterdam professor of mathematics Diederik Johannes Korteweg and his doctoral student Gustav de Vries derived a model equation [42] which describes the unidirectional propagation of long waves in water of relatively shallow depth. This famous equation is now known as the Korteweg–de Vries equation or KdV equation, and it has the following form:

$$\frac{\partial \eta}{\partial t} + V_0 \frac{\partial \eta}{\partial x} + \frac{3}{2}\frac{V_0}{d}\eta\frac{\partial \eta}{\partial x} + \frac{1}{6}V_0 d^2 \frac{\partial^3 \eta}{\partial x^3} = 0\,. \tag{3}$$

Korteweg and de Vries showed that this nonlinear wave equation has a localized solution which represents a single hump of positive elevation corresponding to the solitary wave observed by Scott Russell. The KdV equation had been in fact formulated in 1877 by Boussinesq in an impressive paper [6], but this author did not use directly this formulation. Russell's experiments had found their theory and one could think that many scientists would rapidly extend the results of this theory. However, the KdV equation had a quiet life for many decades.

In the early 1950s, the physicists Enrico Fermi, John Pasta and Stan Ulam could use one of the first computers, the "Maniac" to work on systems without closed analytic solutions. In particular, they studied the way a crystal evolves towards the thermal equilibrium through the numerical simulation of the dynamics of a one-dimensional lattice consisting of N equal masses (atoms) connected by nonlinear springs. The problem is described by the Hamiltonian

$$H = \sum_{i=0}^{N-1} \frac{1}{2} p_i^2 + \sum_{i=0}^{N-1} \frac{1}{2} K (x_{i+1} - x_i)^2 + \frac{K\alpha}{3} \sum_{i=0}^{N-1} (x_{i+1} - x_i)^3 \quad (4)$$

where x_i is the atom i displacement, p_i the corresponding momentum, K the constant of the quadratic potential, and $\alpha (\ll 1)$ a coefficient of nonlinear interaction. The boundary condition $x_0 = x_N = 0$ is assumed. Considering a normal mode decomposition

$$A_k = \sqrt{\frac{2}{N}} \sum_{i=0}^{N-1} x_i \sin(ik\pi/N)\,, \quad (5)$$

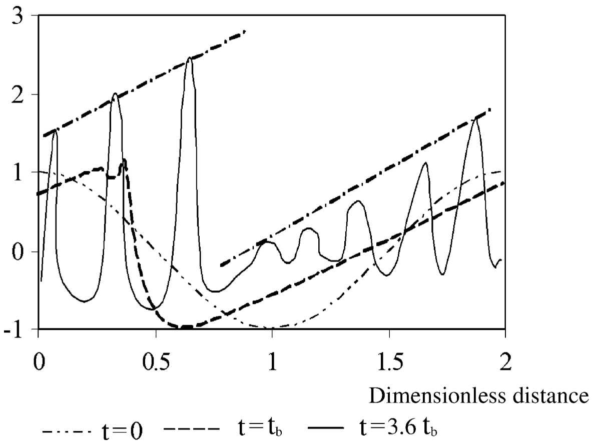
the Hamiltonian (Eq. (4)) may be written in the following way

$$H = \frac{1}{2} \sum_k \left(\dot{A}_k^2 + \omega_k^2 A_k^2 \right) + \alpha \sum_{k,l,m} c_{klm} A_k A_l A_m \quad (6)$$

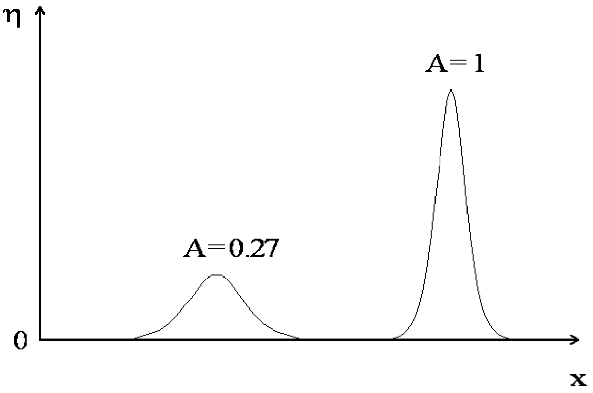
where the pulsations ω_k are given by $\omega_k^2 = 4K \sin^2(k\pi/2N)$, and c_{klm} are constants. When the masses have very small displacements from the equilibrium positions, the dynamics of the lattice may be described by a superposition of normal modes. It was supposed that if only one mode (the fundamental mode $k = 1$ in the circumstances) is excited in a crystal lattice, nonlinear interactions would lead to energy equipartition or thermal equilibrium. This was suggested by the results at the beginning of the calculations since the modes $2, 3, \ldots$ were successively excited. However, after 157 fundamental periods, almost all the energy returned to the fundamental mode. Furthermore, the energy returned almost periodically to the original excited mode leading to a near recurrence process, now known as the FPU paradox. Fermi et al. [18] did not get the expected result and the modeled crystal lattice did not approach energy equipartition. This remarkable discovery is at the origin of the comprehension of the concept of the soliton and of the developments of "numerical experiments". Fermi died just after this work; this had a rather inhibitory effect on the significance of this result and no paper was published after the preprint "Studies of nonlinear problems" [18]. The results were communicated to the scientific community by Ulam through several conferences. Complementary studies [20,39] confirmed that the introduction of nonlinearity in a system does not guarantee the equipartition of energy. In 1960, the KdV equation was rederived by Gardner and Morikawa who worked on collisionless hydromagnetic waves. These authors attracted the attention of Zabrusky and Kruskal on the KdV equation, and the FPU paradox was solved a few years later in terms of solitons [82]. Zabrusky and Kruskal considered the following motion equations corresponding to the Hamiltonian (Eq. (4))

$$\ddot{x}_i = K(x_{i+1} + x_{i-1} - 2x_i) + K\alpha \left[(x_{i+1} - x_i)^2 - (x_i - x_{i-1})^2 \right]. \quad (7)$$

Taking into account that the recurrence occurs before the excitation of high-order modes, these authors carried out an asymptotic analysis in the continuum approximation and obtained an equation including dispersive and nonlinear terms which correspond to the KdV equation. This equation, obtained in a totally different physical context as the one to explain the observations of Scott Russell, is at the origin of the explanation of the FPU paradox. As depicted in Fig. 3, the numerical experiments of Zabrusky and Kruskal showed that the initial sinusoidal condition evolves into steep fronts and then into a finite number of short pulses which are solitons. These solitons travel around the lattice with velocities depending on their amplitude; they collide preserving their individual shapes and velocities with only a small change in their phases. As time increases, there is an instant at which the solitons collide at the same point, and the initial state comes close to recurrence. The recurrence period calculated by Zabrusky and Kruskal closely approximates the actual FPU recurrence period, showing that the KdV equation is a suitable approximation of the FPU system. The system had to be considered as totally nonlinear to be solved, and the linear normal modes of the system could not be considered. The pulse-like waves were called solitons by Zabrusky and



Solitons: Historical and Physical Introduction, Figure 3
Evolution of an initially periodic profile from Zabusky and Kruskal numerical analysis of the KdV equation. The breaking time for the wave profile is t_b



Solitons: Historical and Physical Introduction, Figure 4
Comparison between two solutions of the KdV equation with different amplitude *A*

Kruskal to indicate the remarkable quasi-particle properties of these very stable solitary waves. They were at first called solitrons to introduce them into the family of particle names such as electron and neutron, but the name solitron was a trade mark, and was therefore not suitable. The concept of solitons has spread over numerous branches of physics, such as optical physics, biological physics, astronomy, particle physics, condensed matter, fluid dynamics, ferromagnetism. The numerical experiments of Zabrusky and Kruskal led to the development of integrable systems. Taniuti and Wei [76] and Su and Gardner [74] have shown that a large class of nonlinear evolution equations may be reduced to consideration of the KdV equation which may be regarded as a very important canonical form. A famous method, the so-called Inverse Scattering Transform (IST) was proposed by Gardner et al. [24] to integrate the KdV equation. This procedure may be considered as the nonlinear analogue of the Fourier transform method suitable to linear dispersive systems.

Physical Properties of Solitons and Associated Applications

Properties of Solitons

Let us consider the KdV equation (Eq. (3)). Introducing the variable $X = x - V_0 t$, this equation may be written in the following way

$$\frac{1}{V_0}\frac{\partial \eta}{\partial t} + \frac{3}{2d}\eta\frac{\partial \eta}{\partial X} + \frac{d^2}{6}\frac{\partial^3 \eta}{\partial X^3} = 0\,. \tag{8}$$

Using the dimensionless variables $\phi = \eta/d$, $\xi = X/X_0$ and $\tau = t/t_0$ where X_0 and t_0 are respectively a typical length and time, the KdV equation may be formulated in its standard form:

$$\frac{\partial \phi}{\partial \tau} + 6\phi\frac{\partial \phi}{\partial \xi} + \frac{\partial^3 \phi}{\partial \xi^3} = 0\,. \tag{9}$$

This equation has many solutions, as all nonlinear equations. Among these solutions, there are the spatially localized solutions:

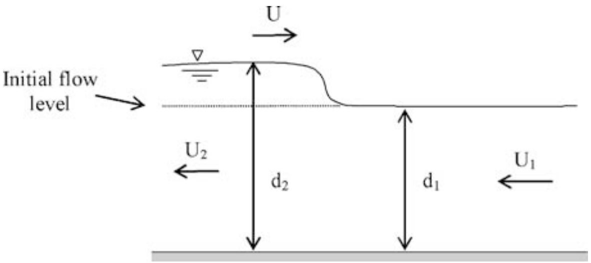
$$\phi = A\,\mathrm{sech}^2\left[\sqrt{\frac{A}{2}}(\xi - 2A\tau)\right], \tag{10}$$

where A is a positive constant. These solutions quantitatively correspond to the observations of Scott Russell. In particular, the width of the soliton decreases for increasing values of its amplitude, as depicted in Fig. 4. In dimensional variables and in a fixed referential, the corresponding solution is given by Eq. (2).

The KdV equation appears widely in physics, each time that waves propagate in a weakly nonlinear and weakly dispersive medium. The nonlinear term $\phi(\partial\phi/\partial\xi)$ in this equation causes the steepness of the wave form. In a dispersive medium, the Fourier components of a pulse propagate at different velocities, inducing a spreading of the pulse. The dispersive term $\partial^3\phi/\partial\xi^3$in the KdV equation makes the wave form spread. The soliton solution of the KdV equation results from the balance of nonlinearity and dispersion. This balance is generally very stable, which explains the numerous applications of the theory of solitons, even if the real physical situations (where friction and defects take place) are only approximately described by the soliton equations, and in particular by the KdV equation. It would be strictly speaking more appropriate to use the terminology quasi-soliton in physics; however, bearing in mind that the word soliton has a quite ideal connotation in the context of systems with exact solutions, we will use the word soliton. The conservation of the shape and of the velocity of the soliton after a collision is a manifestation of stability. The high stability of solitons relative to perturbations, combined with the fact that they can spontaneously emerge in a physical system in which energy is supplied, lead to numerous applications of solitons in macroscopic and microscopic physics. Nevertheless, a soliton may sometimes vanish. For example, this is the case when such a wave propagates in a media where the water depth decreases during the wave propagation. The dispersive effect progressively decreases when the nonlinear effect increases, leading to the wave breaking as waves on a beach. The vanishing of a soliton may also result from energy dissipation. Marin et al. [52] have shown that sand ripples may form when solitons propagate in shallow water over a sandy bed, and that a strong interaction between the free surface and the bed occurs. This interaction leads to a decrease of the soliton amplitude, which may result under certain circumstances in the disappearance of the soliton. Let us now consider some of the most typical soliton applications.

Solitons in Fluid Mechanics

In the hydrodynamic field, a very nice case where solitons take place is the tidal bore, also called hydraulic jump in translation, and known as the mascaret in France. A tidal bore is a positive wave of translation which can occur in a river with a gently slopping bottom and a broad funnel-shaped estuary, where a difference of more than six meters between high and low water takes place [50]. The tidal bore appears as a single wave or a series of waves which propagate upstream. At a given instant, it is localized at the upstream extremity of the flood tide propagation in the inter-tidal zone. The flow direction is towards the sea before the tidal bore arrival, and may be in the opposite direction after its occurrence. Such bores are distributed widely throughout the world. The most powerful can be seen in the Amazon river in South America, where it is called pororoca, which means gigantic noise. An approximately 5 m-high wave propagates with a velocity of about 30 m/s, and surfers take advantage of this exceptional wave to organize competitions. In England, bores may be contemplated on several rivers, the most famous occurring in the Severn river, near Bristol. In France, bores take place mainly in the Mont Saint-Michel bay, in the Gironde, Dordogne and Garonne rivers. The mascaret of the Seine river almost disappeared between 1960 and 1970 because of the construction of dikes and the dredging works which had been carried out in order for large boats to reach the city of Rouen. Increasing the water depth induces a decrease of the nonlinear effects which become insufficient to balance the dispersion. In Canada, several bores are observed in Fundy bay. In Asia, an important bore occurs in the Qiantang river, while the one in the river Pungue in Africa can propagate upstream over more than 80 km. Two types of bores may be observed: the breaking bore and the undulated bore, depending on the value of the Froude number F_r defined by $F_r = (U_1 + U)/\sqrt{gd_1}$, where U_1 is the flow velocity at the depth d_1 before the passage of the bore and U the bore velocity (Fig. 5). In Fig. 5, d_2 depicts the water depth after the passage of the bore and U_2 the flow velocity at the depth d_2. For a value of F_r lower than about 1.5, the bore is undulated, and for a value of F_r greater than 1.5, the breaking bore occurs. Most of the bores are undulated bores. The tidal bore is a very important turbulent phenomenon for the inter-tidal zone of a river, inducing a significant mixing of waters and complex motions of sediment on the river bed [14]. River bores have been known for many centuries but only qualitative observations were carried out. Quantitative measurements came much later, mainly because solitons are inherently nonlinear when only linear processes were considered.



Solitons: Historical and Physical Introduction, Figure 5
Sketch of the propagation of a positive wave of translation

Long waves such as tsunamis often behave like solitary waves [48]. In particular, the run-up and shoreward inundation are often simulated using solitary waves [49]. The tsunamis may be described by the KdV equation. The theory of very long shallow water waves is valid since the tsunamis have a wavelength which is generally greater than 100 km and they propagate in oceans whose mean depth is about 4000 m. Their propagation velocity is greater than $V_0 = \sqrt{gd}$, that is greater than approximately 700 km/h. When they approach a coast, their velocity decreases and their height increases to satisfy the equation of energy conservation, and they finally break inducing often very significant damages on the shore. Most tsunamis have a seismic origin, such as the dramatic one which occurred in the Indian Ocean on 26 December 2004 and which led to the death of more than 220 000 people. These solitons may also result from a quick arrival in the sea of volcanic products, through a similar process to that in Scott Russell's experiments depicted in Fig. 1. This was the case for the tsunami generated in 1883 by the eruption of Krakatau in Indonesia, when approximately 36 000 people died. The inundation phase for a tsunami highly depends on the bottom bathymetry. The breaking can be progressive and structural damages are mainly caused by inundation [13]. The breaking can also be explosive and induce the formation of a plunging jet. When the breaking takes place in very shallow water, the tsunami amplitude becomes so high that an undulation appears on the long wave, which develops into a progressive bore [10]. This turbulent front is of the same type as the wave which occurs when a dam breaks, and the water can rise very rapidly, typically from 0 to 3 meters in 1.5 minutes. The event of December 2004 has shown that the available models are not able to predict accurately the wave run-up heights under severe conditions.

Helal and El-Eissa [37] presented the connection between shallow water waves and the KdV equation. More recently, Mei and Li [55] considered the propagation of long waves over a randomly rough seabed. These authors have shown analytically and numerically that, assuming a randomness height to mean depth ratio comparable to the one of mean depth to the characteristic wavelength, disorder causes diffusion, leading to spatial attenuation of the wave amplitude. Mei and Li [55] proposed a modified KdV equation which includes terms representing the effects of disorder on amplitude attenuation. After a propagation over a region of finite length, the transmitted wave is a pulse which disintegrates into several small solitons of vanishing energy.

Hydrodynamic solitons may also be observed in deep water. Stokes showed in 1847 through the use of a small amplitude expansion of a sinusoidal wave, that periodic waves of finite amplitude are possible in deep water. The Stokes waves are unstable to infinitesimal modulation perturbations [4]. Zakharov and Shabat [83] were the first to show that an initial wave packet may evolve into a number of envelope solitons and a dispersive tail, the envelope solitons consisting of a carrier wave modulated by an envelope signal. This was experimentally verified by Yuen and Lake [81], by carrying out detailed experiments in deep water in a wave tank. The stability properties of envelope solitons were also considered. The freak waves, also called rogue waves, extreme or giant waves, may also have soliton-like shape. These waves appear surprisingly in the sea ("wave from nowhere") in deep or shallow water, and can cause severe damages to ships and fixed ocean structures. They are characterized by their brief occurrence in space and time, resulting from a local focusing of wave energy [19]. The main features of their physical processes may be obtained performing numerical simulations in the framework of classical nonlinear evolution equations, such as the KdV equation [41].

Oceanographers have also observed internal solitary waves in many regions around the world's oceans [61]. Most of the observed solitons propagate at the interface between the thermocline and the deep sea [56]. They are mainly excited by tidal flows over bottom topography [25], and they contribute to significant vertical mixing. Brandt et al. [7] developed a rotationless Boussinesq-like model for the generation and propagation of internal waves; the results of the simulations are in good agreement with large solitary waves recorded with airborne Radar images. Michallet and Barthélemy [56] carried out experiments in a 3-m-long flume to study interfacial long-waves in a two-immiscible-fluid system involving water and petrol. The comparison between the experiments and nonlinear theories shows that the KdV solitary waves match the experiments for small amplitude waves for all layer thickness ratios. When the crest is close to the critical level, that is approximately located at mid-depth, the large amplitude waves are well predicted by a modified KdV equation including both quadratic and cubic nonlinear terms (KdV-mKdV equation). However, only very few laboratory or field measurements of internal solitary waves are available, due to the prohibitive cost of obtaining precise measurements in the field and to the difficulty of generating and measuring these waves in the laboratory.

Blood pressure pulses may be considered as solitons [63]. There is a balance between the nonlinearity due to the blood flow and the dispersion connected to the elasticity of the artery. Previous studies [32,33,80] have shown that the dynamics of weakly nonlinear pressure waves in a thin nonlinear elastic tube filled with an incompressible fluid may be governed by the KdV equation. A Boussinesq-like equation was obtained by Paquerot and Remoissenet [62] to describe the blood pressure propagation in large arteries, in the limit of an ideal fluid and for slowly varying arterial parameters.

Several methods have been tested to generate hydrodynamic solitons in a flume; the method used by Scott Russell has been described in Sect. "Introduction" (Fig. 1). Hammack and Segur [29] showed experimentally and theoretically that from any net positive volume of water above the still water level, at least one solitary wave will emerge followed by a train of dispersive waves. Maxworthy [54] and Weidman and Maxworthy [79] pushed the liquid in the horizontal direction by a single displacement of a piston. The generation of solitary waves using a piston-type wave maker has been studied in detail by Goring [26]. Guizien and Barthélemy [28] have proposed an experimental procedure to generate solitary waves in a flume using a piston type wave maker from Rayleigh's (1876) solitary wave solution. The advantage of this method is that solitary waves are rapidly established with very little loss of amplitude in the initial stage of the propagation of the solitary waves. The generation of solitons using a wave flume in resonant mode, without an absorbing beach, has been considered by Marin et al. [52]. Surface waves are produced in shallow water by an oscillating paddle at one end of the flume; a near-perfect reflection takes place at the other end. The frequency of the oscillating paddle was chosen close to the resonant frequency of the mode whose wavelength is equal to the flume length. For small values of the amplitude of displacement of the oscillating paddle, only standing harmonic waves are generated in the flume. For values of this amplitude greater than a critical value, pulses propagating from one end of the channel to the other end are excited on the background of the standing harmonic wave. From one to four pulses propagating in each direction of the flume could be generated over the time period of the flow, depending on the frequency and the amplitude of horizontal displacement of the oscillating paddle. These pulses were identified as solitons (one pulse) and bound states of solitons (several pulses). The generation of solitons in the flume results from the excitation of high harmonics. The spatiotemporal properties of solitons generated in this way were studied in detail by Ezersky et al. [17]. Space-time diagrams have been constructed to highlight the spatiotemporal dynamics of nonlinear fields for two solitons colliding in the resonator. Period doublings and the multistability of the nonlinear waves, i. e. the occurrence of different regimes for the same values of the control parameters but under different initial conditions, have been shown. It is important to keep in mind that the solitary waves correspond to the limit of cnoidal waves of period tending to infinity. The velocity of cnoidal waves V_{cn} is known to depend on the so-called elliptic parameter m and is given by the formula [11]

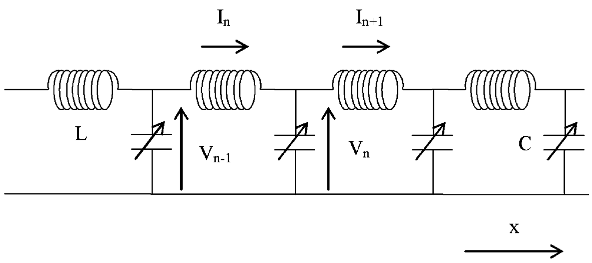
$$V_{\mathrm{cn}} = \left[1 - \frac{A_{\mathrm{S}}}{2d} + \frac{A_{\mathrm{S}}}{md}\left(1 - \frac{3E(m)}{2K(m)}\right)\right]\sqrt{gd} \qquad (11)$$

where $K(m)$ and $E(m)$ are complete elliptic functions of the first and second kind. The parameter m, which depends on the wave amplitude to width ratio, is responsible for the shape of the wave with $m = 0$ corresponding to the harmonic wave, and $m = 1$ to the soliton. Strictly speaking, the waves generated in a flume used in resonant mode are not exactly solitons, but cnoidal waves with an elliptic parameter very close to 1; the value of this parameter was found to be 0.9996 for the tests carried out by Ezersky et al. [17]. The shapes of pulses are almost indistinguishable for $m = 1$ and $m = 0.9996$, but the pulse repetition rates differ appreciably, the period tending to infinity as $m \to 1$.

Atmospheric solitons also exist, such as the Morning Glory Cloud of the Gulf of Carpentaria in northern Australia, which can be seen as a spectacular propagating roll cloud. Atmospheric solitons are horizontally propagating nonlinear internal gravity waves that can travel over large distances with minimal change in form [68]. The solitary waves that have been observed in the atmosphere can be divided into two classes: those that propagate in a fairly shallow stratified layer near the ground and those that occupy the entire troposphere. The generation mechanisms appear to be quite different for these two classes of waves. The waves that occupy the lower part of the troposphere mainly involve a gravity current such as a thunderstorm outflow, a katabatic wind, a sea breeze front, or a downslope windstorm, interacting with a low-level stable layer. The motion of the gravity current produces perturbations that are trapped in the low-level stable layer and eventually evolve over time into a series of solitary waves. For the tropospheric scale waves, they are very probably generated by synoptic scale features such as large-scale convective systems and geostrophic adjustments. For the two types of atmospheric solitons, there must exist some feature in the atmosphere that serves to prevent the wave energy from propagating away in the vertical direction. These trapping mechanisms are either deep layers of the atmosphere with very low values of the buoyancy frequency, or critical layers. The KdV equation, eventually enhanced with higher-order nonlinearity, gives a reasonable agreement with the observations [68].

Solitons in Nonlinear Transmission Lines

In another area of physics, transmission lines with nonlinear elements are found to propagate in a soliton mode. Nonlinear electrical transmission lines are simple experimental devices to observe and consider quantitatively the propagation and properties of solitons. The propagation of these waves is used for picosecond impulse generation and broadband millimeter-wave frequency multiplication [66]. Let us consider the elementary LC network depicted in Fig. 6, with linear inductors L and nonlinear capacitors C. The differential capacitance $C(V_n)$ is supposed to depend nonlinearly on the voltage across the nth



Solitons: Historical and Physical Introduction, Figure 6
Elementary circuit of an electrical network with linear inductance's *L* and nonlinear capacitance's *C*

capacitor V_n:

$$C(V_n) = \frac{\mathrm{d}Q_n(V_n)}{\mathrm{d}V_n} \tag{12}$$

where $Q_n(V_n)$ is the charge stored in the nth capacitor. Following Kirchhoff's law, we have:

$$V_{n-1} - V_n = \frac{\mathrm{d}\phi_n}{\mathrm{d}t}\,, \quad I_n - I_{n+1} = \frac{\mathrm{d}Q_n}{\mathrm{d}t} \tag{13}$$

where the magnetic flux ϕ_n is related to the current I_n by the relation: $\phi_n = LI_n$. Using Eqs. (12) and (13), it is easy to obtain:

$$\frac{\mathrm{d}^2 Q_n}{\mathrm{d}t^2} = \frac{1}{L}(V_{n+1} + V_{n-1} - 2V_n)\,, \quad n = 1, 2, \ldots \tag{14}$$

Assuming the following capacitance-voltage relation,

$$C(V_n) = C_0(1 - 2aV_n) \tag{15}$$

where a is a small nonlinear coefficient, we find:

$$\begin{aligned} &LC_0\frac{\mathrm{d}^2 V_n}{\mathrm{d}t^2} - LC_0 a\frac{\mathrm{d}^2 V_n^2}{\mathrm{d}t^2} \\ &\qquad = (V_{n+1} + V_{n-1} - 2V_n)\,, \quad n = 1, 2, \ldots \end{aligned} \tag{16}$$

The system (Eq. 16) of nonlinear equations cannot be solved analytically. The continuum limit is used to get approximate solutions; setting the position $x = n\delta$ where δ is a hypothetical cellule length, we get

$$\frac{\partial^2 V}{\partial t^2} - \frac{\delta^2}{LC_0}\frac{\partial^2 V}{\partial x^2} = \frac{\delta^4}{12LC_0}\frac{\partial V}{\partial x^4} + a\frac{\partial^2 V^2}{\partial t^2}\,. \tag{17}$$

This weakly dispersive and nonlinear wave equation describes waves that can travel both to the left and to the right; the dispersive nature is due to the discreteness of the electrical network. A localized wave solution of the Eq. (17) that does not change its shape as it propagates with constant velocity v may be found [65]:

$$V = \frac{3(v^2 - v_0^2)}{2av^2}\operatorname{sech}^2\left(\frac{\sqrt{3(v^2 - v_0^2)}}{v_0}\left(n - \frac{v}{\delta}t\right)\right) \tag{18}$$

where $v_0 = \delta/\sqrt{LC}$. This solitary wave solution represents a pulse with amplitude

$$V_m = \frac{3}{2a}\frac{v^2 - v_0^2}{v^2} \tag{19}$$

which depends on the velocity v. The width L_w at half height of this pulse is given by

$$L_w = 1.76\frac{v_0}{\sqrt{3(v^2 - v_0^2)}}\,. \tag{20}$$

The waveforms of this solitary wave and of the KdV soliton are similar, and the width L_w depends on the amplitude V_m:

$$V_m L_w^2 = C^t \tag{21}$$

where C^t is a constant.

Solitons in Plasmas

Solitons may propagate in plasmas; this has received much attention because of a possible relevance to final state configurations in fusion devices. The combination of dispersion and nonlinearity in plasmas may be described by the KdV equation. A direct analogy between the equations for shallow water and those for plasmas has been mentioned by Gardner and Morikawa [23]. However, the KdV equation is not always adapted to the excitations which can be generated in plasmas by electromagnetic waves, since wave packets are often produced instead of pulses characterizing the solutions of the KdV equation. Another type of soliton equation is then more suitable, the nonlinear Schrödinger equation (NLS); this second class of soliton equation will be considered in Subsect. "Solitons in Optical Fibers".

Solitons in a Chain of Pendulums

Hydrodynamic solitons, solitons in nonlinear transmission lines are nontopological solitons, because the system returns to its initial state after the passage of the wave. There is also another type of soliton, the kink solitons which are topological solitons, since the structure of the system is sometimes modified after the passage of the wave. A mechanical system consisting of a chain of pendulums, each pendulum being elastically connected to its neighbors by springs, is a typical example for which topological solitons can merge. It is easy to show that this mechanical system may be described by the following equation:

$$\frac{\partial^2\theta}{\partial t^2} - c_0^2\frac{\partial^2\theta}{\partial x^2} + \omega_0^2 \sin\theta = 0 \tag{22}$$

where θ is the angle of rotation of the pendulums, x is the axis along which the pendulums are distributed, $c_0^2 = l_1^2\beta/I_i$, l_1 being the distance between two pendulums, β the torque constant of a section of spring between two pendulums, I_i the moment of inertia of a single pendulum of mass m and length l_2, and $\omega_0^2 = mgl_2/I_i$. The first term in Eq. (22) represents the inertial effects of the pendulums, the second term corresponds to the restoring torque due to the coupling between pendulums, and the third term represents the gravitational torque. The Eq. (22) is called the Sine–Gordon (SG) equation [47]; it can be totally integrated and admits exact solitons solutions. It has become a very famous equation containing dispersion and nonlinearity which is used to model various phenomena in physics. It constitutes the third class of nonlinear equation leading to solitons.

Fluxons in a Josephson Tunnel Junction

Solitons arise also in a Josephson tunnel junction, which is a junction between two superconductors. These junctions are very attractive for information processing and storage [78]. The physical quantity of interest in the long Josephson junction is a quantum of magnetic flux, or a fluxon, which has a soliton behavior. The relevant nonlinear equation which describes the physical processes is the Sine–Gordon equation. This equation is particularly suitable in solid physics.

Solitons in Optical Fibers

One of the fundamental applications of solitons concerns optical fibers. In optical fiber communication using linear pulses, dispersion and losses in the fiber limit the information capacity which can be transported and the distance of transmission. Hasegawa and Tappert [31] first proposed to balance the dispersive effect by nonlinearity using soliton-based communications. The Kerr effect, in other words the nonlinear change of the refractive index of the fiber material, is used and an initial optical pulse may form a nonlinear stable pulse, an optical soliton, which is in fact an envelope soliton. Hasegawa and Tappert proposed the previously mentioned nonlinear Schrödinger equation for the description of solitons propagating in optical fibers. The first observations of these solitons were made in 1980 by Mollenauer, Stolen, and Gordon [58]. In real media, the light intensity of the soliton decreases due to losses in the fiber, and suitable reshaping of the pulse is required. Numerous theoretical and experimental studies on nonlinear guided waves have been carried out, because of their wide range of applications [30,43]. Inserting an optical fiber in a laser cavity, a "soliton laser" may be obtained, a femtosecond laser which is now in use. A real revolution in international communications is taking place at present, based on optical soliton technology. It is anecdotal that a fiber-optic cable linking Edinburgh and Glasgow now runs beneath the path from which John Scott Russell made his initial observations, and along the aqueduct which now bears his name.

The NLS equation may be obtained from the Sine–Gordon equation. Previously considered systems as water waves on the free surface or chains of pendulums have also weak amplitude solutions, plane waves which have the form:

$$\theta = \psi\, e^{i(qx-\omega t)} + \text{c.c.} \tag{23}$$

where ψ is the wave amplitude, q is the wave number, ω is the pulsation and c.c. denotes the complex conjugate of a complex number. When the wave amplitude is sufficiently increased for the nonlinearity to take place, modulation can spontaneously arise because harmonics are generated by the nonlinearity. The initial wave may split in waves packets whose properties correspond to the ones of solitons. These solitons consisting of a carrier wave which is modulated by an envelope signal are called envelope solitons. It can be shown [63] that the evolution of the envelope ψ is described by the NLS equation:

$$i\frac{\partial\psi}{\partial t} + P\frac{\partial^2\psi}{\partial x^2} + Q\,|\psi|^2\,\psi = 0 \tag{24}$$

where P and Q are coefficients which depend on the physical problem, as the significance of the variables t and x. The NLS equation is formally similar to the Schrödinger equation of quantum mechanics:

$$i\hbar\frac{\partial\psi}{\partial t} + \frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2} - U\psi = 0\,. \tag{25}$$

The potential U in Eq. (25) is proportional to the absolute square of the wave envelope ψ, m is the mass of the quasi-particle and $\hbar = h/2\pi$ where h is the Planck constant. The potential represents the self-trapping of the wave energy.

Solitons in Solid Physics

Solitons are involved in the atomic structure of matter. At the beginning of the twentieth century, a central question in solid state physics was how a plastic deformation of a metal takes place in the corresponding crystal lattice. This problem induced numerous discussions, and it was suggested that if one atom in the lattice is pulled away, a neighbor atom follows the one which had been moved, jumping into the new hole, that is the new free lattice space. Then, a dislocation runs through the crystal. This proposition was improved in 1939 by Frenkel and Kontorova [22] who developed a model where not only each atom moves alone over one lattice distance, but also many atoms form a dislocation line generating a wave of translation. The process of traveling of the dislocation line through the crystal was assumed to occur without losses of energy. This is equivalent to Russell's solitary wave of translation which travels over long distance without change of form and velocity. This led Frenkel and Kontorova to make an analogy between the dislocation line and the soliton, and they obtained the one-soliton solution of the Sine–Gordon equation.

The solitons permit us to interpret properties of dielectric materials. In other respects, magnetic materials are interesting examples to experimentally verify in a very accurate way the theory of solitons at the atomic scale [57]. The concept of soliton in polymer physics constitutes a very nice case of an interdisciplinary approach. Their occurrence was suggested in 1988 by theoretician physicists [34]. Many studies have since then been carried out, in chemistry and experimental physics. Alan J. Heeger, Alan G. MacDiarmid and Hideki Shirakawa received the Nobel prize in chemistry in October 2000 for their works on the electric conduction of conductive plastics which is performed by solitons.

Solitons in Biology

Interest in the physical modeling of biological processes has grown significantly since the beginning of the 1990s. Nonlinear localization phenomena have been recently shown in systems very close to biological systems [15]. The notion of the soliton is now used to explain the dynamics of biological macromolecules such as proteins and the molecule ADN; an approach to the dynamics of the molecule ADN can be obtained using the envelope solitons solution of the NLS equation [63]. Intense researches are presently being carried out on this subject. The great size of biological molecules allows collective behaviors of atom groups, which are with nonlinearity important components for the existence of solitons.

Mathematical Methods Suitable for the Study of Solitons

Solitons may appear in many fields, such as fluid mechanics, solid state physics, plasma physics, optical fibers and biology, as shown in Sect. "Physical Properties of Solitons and Associated Applications". The description of these physical systems often leads to dispersive and nonlinear equations which are not presently solvable. It is then important to refer to the great classes of soliton equations, which are idealized models for numerous systems. These great classes of soliton equations correspond to the Korteweg–de Vries equation, known as the KdV equation (Eq. (9)), the Sine–Gordon equation (SG; Eq. (22)), and the nonlinear Schrödinger equation (NLS, Eq. (24)). These three famous equations can be totally integrated. They are not completely dissociated. For instance, it is possible to obtain the NLS equation from the SG equation and from the KdV equation. The soliton solutions represent only a first approximation of the physical properties of real systems. Dissipation or a weak spatial or temporal variation of the physical parameters in real physical systems are examples of phenomena which are very common and which are not taken into account in the soliton equations. The modeling has to be adapted to the physical system and to the excitation conditions of this system. In the case of plasmas, the KdV equation is more suitable for a description of the system when intense pulses excite it, and the NLS equation is more adapted for sinusoidal perturbations.

There are two broad theoretical approaches available to obtain the solutions of the great classes of soliton equations, the analytic and the algebraic methods. Analytic approaches include the powerful inverse scattering transform (IST) [24] and the remarkable Hirota method [38]. One-soliton solutions or multi-soliton solutions can be obtained. Moreover, the IST method can predict the number of solitons that can emerge from an initial disturbance applied to a physical system [59]. This method shows how the solitons have a role in nonlinear normal modes as the Fourier modes for a linear equation [44]. There are also the Bäcklund transformation method [67], the direct linearizing transform method [1], the Painlevé analysis [45], and the method of position-like solutions [40]. The algebraic methods involve the Lie group theoretic method [60], the direct algebraic method [53], and the tangent hyperbolic method [51].

The theoretical methods cannot always be used since they need some conditions to make them applicable. Numerical methods can be applied [36]. Zabusky and Kruskal [82] were the pioneers in studying in 1965 the KdV equation numerically, by using the leapfrog method as an explicit finite difference scheme. Many methods have since then been used: a split-step Fourier method [77], the hopscotch method [27], a pseudo-spectral method [21], spectral methods (the Galerkin and Chebyshev methods) [35,72], the finite element method [75], and Fourier collocation calculations [9].

Future Directions

Solitons may occur in the macroscopic and microscopic scales of physics. This results from their very general existence conditions, the coexistence of dispersion and nonlinearity. The concept of solitons is not restricted to a specific field of physics, and the investigation of solitons remains a very active research area. We have previously mentioned (Sect. "Physical Properties of Solitons and Associated Applications") that one important application of solitons concerns optic fibers. The propagation of optical solitons through nonlinear fibers has been extensively studied; however, in real media, the dynamics of these solitons and the conditions for their generation are significantly affected by various inhomogeneities in the media. The problem of nonlinear wave propagation in the form of solitons in inhomogeneous optical fiber media is actually not well understood, despite the wide range of applications [30]. The investigation of nonlinear processes in coupled optical waveguides is another research direction for the future design of optical computers and sensor elements. Recent advances have shown that solitons have a great potential for the improvement of optical systems which demand fast and reliable data transfer. In the hydrodynamic field, the tsunamis often behave like solitary waves. The dramatic tsunami which occurred in the Indian Ocean on 26 December 2004 has clearly shown that the actual numerical models do not accurately predict the water propagation on the coastal plains. An important direction for future research is the (3D) three-dimensional simulation of tsunami breaking along a coast. The 3D simulation of the interaction of a tsunami with different beach profiles, with and without obstacles has also to be considered. This will permit the design of protecting devices and the setting up of security zones. In biology, nonlinear localization phenomena have been proved in model systems close to biological systems, but the challenge is open for biological molecules. Other important application fields for solitons concern magneto-electronics and secure communications. The formation of spatio-temporal patterns on perturbations of soliton systems are important problems to be studied in the years to come. The existence of instabilities in the solitary wave solutions of the great classes of soliton equations have to be tackled. Nonlinear excitations in systems whose spatial dimension is greater than one raise many questions. Among these questions, there are the problems of the possible stable structures, their collision properties and the effect of external forces. New applications will certainly emerge from these studies.

Bibliography

Primary Literature

1. Ablowitz MJ, Clarkson PA (1991) Solitons, nonlinear evolutions equations and inverse scattering. Cambridge University Press, Cambridge
2. Airy GB (1845) Tides and waves. Encycl Metropolitana 5:291–396
3. Bazin H (1865) Recherches expérimentales relatives aux remous et à la propagation des ondes. In: Darcy H and Bazin H (eds) Recherches hydrauliques, imprimerie impériale, Paris
4. Benjamin TB, Feir JE (1967) The disintegration of wavetrains on deep water. Part 1 Theory. J Fluid Mech 27:417–430
5. Boussinesq J (1871) Théorie de l'intumescence liquide appelée onde solitaire ou de translation, se propageant dans un canal rectangulaire. CR Acad Sci 72:755–759
6. Boussinesq J (1877) Essai sur la théorie des eaux courantes. MSE 23:1–680
7. Brandt P, Rubino A, Alpers W, Backhaus JO (1997) Internal waves in the Strait of Messina studied by a numerical model and synthetic aperture radar images from the ERS 1/2 satellites. J Phys Oceanogr 27:648–663
8. Bullough RK (1988) The Wave Par Excellence, the solitary progressive great wave of equilibrium of the fluid: an early history of the solitary wave. In: Lakshmanan M (ed) Solitons: Introduction and applications. Springer Ser Nonlinear Dyn. Springer, New York, pp 150–281
9. Canuto C, Hussaini MY, Quarteroni A, Zang TA (1988) Spectral methods in fluid dynamics. Springer, Berlin
10. Chanson H (2005) Le tsunami du 26 décembre 2004: un phénomène hydraulique d'ampleur internationale. Premiers constats. Houille Blanche 2:25–32
11. Cooker MJ, Weidman PD, Bale DS (1997) Reflection of high amplitude wave at a vertical wall. J Fluid Mech 342:141–158
12. Darrigol O (2003) The spirited horse, the engineer, and the mathematician: water waves in nineteenth-century hydrodynamics. Arch Hist Exact Sci 58:21–95
13. Dias F, Dutykh D (2007) Dynamics of tsunami waves. Extreme Man-Made and Natural Hazards in Dynamics of Structures. In: Ibrahimbegovic A, Kozar I (eds) Proc NATO Adv Res Workshop on Extreme Man-Made and Natural Hazards in Dynamics of Structures. Springer, Opatija, Croatia
14. Donnelly C, Chanson H (2002) Environmental impact of a tidal bore on tropical rivers. In: Proc 5th Int River Management Symp. Brisbane, Australia
15. Edler J, Hamm P (2002) Self-trapping of the amide I band in a peptide model crystal. J Chem Phys 117:2415–2424
16. Emerson GS (1977) John Scott Russell: a great Victorian engineer and naval architect. John Murray, London
17. Ezersky AB, Polukhina OE, Brossard J, Marin F, Mutabazi I (2006) Spatiotemporal properties of solitons excited on the surface of shallow water in a hydrodynamic resonator. Phys Fluids 18:067104

18. Fermi E, Pasta J, Ulam S (1955) Studies of nonlinear problems. Los Alamos report, LA-1940. published later In: Segré E (ed)(1965) Collected Papers of Enrico Fermi. University of Chicago Press
19. Fochesato C, Grilli S, Dias F (2007) Numerical modelling of extreme rogue waves generated by directional energy focusing. Wave Motion 44:395–416
20. Ford JJ (1961) Equipartition of energy for nonlinear systems. J Math Phys 2:387–393
21. Fornberg B, Whitham GB (1978) A numerical and theoretical study of certain nonlinear wave phenomena. Philos Trans R Soc London 289:373–404
22. Frenkel J, Kontorova T (1939) On the theory of plastic deformation and twinning. J Phys 1:137–149
23. Gardner CS, Morikawa GK (1960) Similarity in the asymptotic behaviour of collision-free hydromagnetic waves and water waves. Technical Report NYO-9082, Courant Institute of Mathematical Sciences. New York University, New York
24. Gardner CS, Green JM, Kruskal MD, Miura RM (1967) Method for solving the Korteweg-de Vries equation. Phys Rev Lett 19:1095–1097
25. Gerkema T, Zimmerman JTF (1994) Generation of nonlinear internal tides and solitary waves. J Phys Oceanogr 25:1081–1094
26. Goring DG (1978) Tsunamis – The propagation of long waves onto a shelf. Ph D thesis. California Inst Techn, Pasadena, California
27. Greig IS, Morris JL (1976) A hopscotch method for the KdV equation. J Comput Phys 20:64–80
28. Guizien K, Barthélemy E (2002) Accuracy of solitary wave generation by a piston wave maker. J Hydraulic Res 40(3):321–331
29. Hammack JL, Segur H (1974) The Korteweg-de Vries equation and water waves. Part 2. Comparisons with experiments. J Fluid Mech 65:289–314
30. Hao R, Li L, Li ZH, Xue W, Zhou GS (2004) A new approach to exact soliton solutions and soliton interaction for the nonlinear Schrödinger equation with variable coefficients. Opt Commun 236:79–86
31. Hasegawa A, Tappert F (1973) Transmission of stationary nonlinear optical pulses in dispersive dielectric fiber: II. Normal dispersion. Appl Phys Lett 23:171–172
32. Hashizume Y (1985) Nonlinear pressure waves in a fluid-filled elastic tube. J Phys Soc Japan 54:3305–3312
33. Hashizume Y (1988) Nonlinear pressure wave propagation in arteries. J Phys Soc Japan 57:4160–4168
34. Heeger AJ, Kivelson S, Schrieffer JR, Su WP (1988) Solitons in conducting polymers. Rev Modern Phys 60:781–850
35. Helal MA (2001) Chebyshev spectral method for solving KdV equation with hydrodynamical application. Chaos Solit Fractals 12:943–950
36. Helal MA (2002) Review: Soliton solution of some nonlinear partial differential equations and its applications in fluid mechanics. Chaos, Solitons and Fractals 13:1917–1929
37. Helal MA, El-Eissa HN (1996) Shallow water waves and KdV equation (oceanographic application). PUMA 7(3–4):263–282
38. Hirota R (1971) Exact solution of the Korteweg–de Vries equation for multiple collisions of solitons. Phys Rev Lett 27:1192–1194
39. Jackson EA (1963) Nonlinear coupled oscillators. I. Perturbation theory: ergodic problems. J Math Phys 4:551–558
40. Jaworski M, Zagrodzinski J (1995) Position and position-like solution of KdV and Sine–Gordon equations. Chaos Solit Fractals 5(12):2229–2233
41. Kharif C, Pelinovsky E (2003) Physical mechanisms of the rogue wave phenomenon. Eur J Mech B/Fluids 22:603–634
42. Korteweg DJ, De Vries G (1895) On the change of form of long waves advancing in a rectangular channel, and on a new type of long stationary waves. Phil Mag 39(5):442–443
43. Kruglov VI, Peacok AC, Harvey JD (2003) Exact Self–Similar Solutions of the Generalized Nonlinear Schrödinger Equation with Distributed Coefficients. Phys Rev Lett 90(11):113902 http://prola.aps.org/abstract/PRL/v90/i11/e113902
44. Lakshmanan M (1997) Nonlinear physics: integrability, chaos and beyond. J Franklin Inst 334B(5/6):909–969
45. Lakshmanan M, Sahadevan R (1993) Painlevé analysis, Lie symmetries and integrability of coupled nonlinear oscillators of polynomial type. Phys Rep 224:1–93
46. Lamb H (1879) Treatise on the motion of fluids. Hydrodynamics, 6th edn 1952. Cambridge University Press, Cambridge
47. Lamb H (1971) Analytical descriptions of ultrashort optical pulse propagation in a resonant medium. Rev Mod Phys 43:99–124
48. Liu PL-F, Synolakis CE, Yeh HH (1991) Report on the International Workshop on long-wave run-up. J Fluid Mech 229:675–688
49. Lo EYM, Shao S (2002) Simulation of near-shore solitary wave mechanisms by an incompressible SPH method. Appl Ocean Res 24:275–286
50. Lynch DK (1982) Tidal bores. Sci Am 247:134–143
51. Malfliet W (1992) Solitary wave solutions of nonlinear wave equations. Am J Phys 60(7):650–654
52. Marin F, Abcha N, Brossard J, Ezersky AB (2005) Laboratory study of sand bedforms induced by solitary waves in shallow water. J Geophys Res 110(F4):F04S17
53. Martnez Alonso L, Olmedilla Morino E (1995) Algebraic geometry and soliton dynamics. Chaos Solit Fractals 5(12):2213–2227
54. Maxworthy T (1976) Experiments on collision between solitary waves. J Fluid Mech 76:177–185
55. Mei CC, Li Y (2004) Evolution of solitons over a randomly rough seabed. Phys Rev E 70:016302
56. Michallet H, Barthélemy E (1998) Experimental study of interfacial solitary waves. J Fluid Mech 366:159–177
57. Mireska HJ, Steiner M (1991) Solitary excitations in one-dimensional magnets. Adv Phys 40:196–356
58. Mollenauer LF, Stolen RH, Gordon JP (1980) Experimental observation of picosecond pulse narrowing and solitons in optical fibers. Phys Rev Lett 45:1095–1098
59. Newell AC (1985) Solitons in mathematics and physics. SIAM, Philadelphia
60. Olver PJ (1986) Applications of Lie groups to differential equations. Graduate Texts in Mathematics, vol 107. Springer, Berlin
61. Ostrovsky LA, Stepanyants YA (1989) Do internal solitons exist in the ocean? Rev Geophys 27:293–310
62. Paquerot JF, Remoissenet M (1994) Dynamics of nonlinear blood pressure waves in large arteries. Phys Lett A 194:77–82
63. Peyrard M, Dauxois T (2004) Physique des solitons. EDP Sciences. CNRS Editions, Paris
64. Rayleigh L (1876) On waves. Phil Mag 5(1):257–279
65. Remoissenet M (1999) Waves called solitons – Concepts and experiments. Springer, Berlin

66. Rodwell MJ, Allen ST, Yu RY, Case MG, Bhattacharya U, Reddy M, Carman E, Kamegawa M, Konishi Y, Pusl J, Pullela R (1994) Active and nonlinear wave propagation devices in ultrafast electronics and optoelectronics. Proc IEEE 82:1037–1058
67. Rogers C, Shadwick WF (1982) Bäcklund transformations and applications. Academic, New York
68. Rottman JW, Grimshaw R (2001) Atmospheric internal solitary waves. In: Environmental stratified flows. Kluwer, Boston, pp 61–88
69. Russell JS (1837) Report on the committee on waves. In: Murray J (ed) Bristol, Brit Ass Rep, London, pp 417–496
70. Russell JS (1839) Experimental researches into the laws of certain hydrodynamical phenomena that accompany the motion of floating bodies, and have not previously been reduced into conformity with the known laws of the resistance of fluids. Trans Royal Soc Edinb 14:47–109
71. Russell JS (1844) Report on waves. In: Murray J (ed) Brit Ass Rep Adv Sci 14, London, pp 311–390
72. Sanz–Serna JM, Christie I (1981) Petrov-Galerkin method for nonlinear dispersive waves. J Comput Phys 39:94–102
73. Stokes GS (1847) On the theory of oscillatory waves. Trans Cambridge Phil Soc 8:441–473
74. Su CH, Gardner CS (1969) Korteweg-de Vries equation and generalizations. III. Derivation of the Korteweg-de Vries equation and Burgers equation. J Math Phys 10:536–539
75. Taha TR, Ablowitz MJ (1984) Analytical and numerical aspects of certain nonlinear evolution equations, (III) numerical, KdV equation. J Comput Phys 55:231–253
76. Taniuti T, Wei CC (1968) J Phys Soc Jpn 24:941–946
77. Tappert F (1974) Numerical solution of the KdV equation and its generalisation by split-step Fourier method. Lec Appl Math Am Math Soc 15:215–216
78. Ustinov AV (1998) Solitons in Josephson junctions. Phys D 123:315–329
79. Weidman PD, Maxworthy T (1978) Experiments on strong interaction between solitary waves. J Fluid Mech 85:417–431
80. Yomosa S (1987) Solitary waves in large blood vessels. J Phys Soc Japan 56:506–520
81. Yuen HC, Lake BM (1975) Nonlinear deep water waves: theory and experiments. Phys Fluids 18:956–960
82. Zabusky NJ, Kruskal MD (1965) Interaction of solitons in a collisionless plasma and the recurrence of initial states. Phys Rev Lett 15:240–243
83. Zakharov VE, Shabat AB (1972) Exact theory of two-dimensional self-focusing and onedimensional self-modulation of waves in nonlinear media. Sov Phys JETP 34:62–69

Books and Reviews

Agrawal GP (2001) Nonlinear Fiber Optics. Academic Press, Elsevier

Akmediev NN, Ankiewicz A (1997) Solitons, Nonlinear Pulses and Beams. Chapman and Hall, London

Braun OM, Kivshar YS (2004) The Frenkel–Kontorova Model. Concepts, Methods and Applications. Springer, Berlin

Bullough RK, Caudrey P (1980) Solitons. Springer, Heidelberg

Davydov AS (1985) Solitons in Molecular Systems. Reidel, Dordrecht

Dodd RK, Eilbeck JC, Gibbon JD, Morris HC (1982) Solitons and Nonlinear Wave Equations. Academic Press, London

Drazin PG, Johnson RS (1993) Solitons: an Introduction. Cambridge University Press, Cambridge

Eilenberger G (1981) Solitons: Mathematical Methods for Physicists. Springer, Berlin

Hasegawa A (1989) Optical Solitons in Fibers. Springer, Heidelberg

Infeld E, Rowlands G (2000) Nonlinear waves, Solitons and Chaos. Cambridge University Press, Cambridge

Lamb GL (1980) Elements of Soliton Theory. Wiley, New York

Toda M (1978) Theory of Nonlinear Lattices. Springer, Berlin

Solitons Interactions

TARMO SOOMERE
Center for Nonlinear Studies, Institute of Cybernetics,
Tallinn University of Technology, Tallinn, Estonia

Article Outline

Glossary

Solitary wave A localized structure which travels with constant speed and shape in otherwise homogeneous medium.

Soliton In the classical nomenclature, a soliton is a nonlinear spatially or temporally localized entity (solitary wave, impulse, wave packet, kink, etc.) of permanent form that retains its identity and shape in interactions with other similar entities. Also frequently used term for solutions of the relevant nonlinear partial differential equations approximately describing such physical entities.

Soliton (colloquially as well as in certain domains) Used for denoting any long-lived localized structures (e. g. solitary waves, self-trapped optical beams, localized vortices) which exhibit low energy radiation and approximately keep their shape.

Soliton solution A solution, usually in closed form, of the nonlinear partial differential equations that exhibit properties of a single soliton. A multi-soliton solution exhibits properties and collective behavior of groups of solitons.

Elastic interaction A generalization of the concept of elastic collisions of ideal mechanical bodies: interaction of localized entities in which the dynamically significant properties of the counterparts such as the total energy, linear and angular momentum, and topological charge or spin are conserved. For low-dimensional and scalar solitons denotes the case in which all the properties of interacting solitons are completely restored after interaction.

Inelastic interaction Interaction of soliton-like structures that modifies one or more dynamically significant properties of the counterparts.

Phase shift A change of the position of a soliton owing to interactions with other solitons compared to the location in which it would be in the absence of other solitons.

Resonant interaction A specific form of soliton interactions leading to the fusion of two or more solitons into a new soliton. Generally associated with an infinite phase shift of the counterparts.

Definition of the Subject

The concept of solitons reflects one of the most important developments in science of the 20th century: the nonlinear description of the world. It is customary to start an introduction to solitons by recalling that it is not easy to give a comprehensive and precise definition of a soliton. Frequently, a soliton is explained as a spatially localized (solitary) wave with spectacular stability properties. Although the combination of simultaneously being a localized structure propagating while mostly keeping its shape as waves do, and surviving for a long time in realistic (that is, nonlinear) conditions, already is a fascinating property in itself, yet another key quality defines whether a particular entity is a soliton. The distinction is made based on the way in which two or more of these objects interact with each other.

The classical nomenclature associates the term soliton with (i) nonlinear and (ii) localized entities, which (iii) have a permanent form and (iv) retain their identity in interactions with other entities from the same class (e. g. Drazin and Johnson [37]; also the entry ▶ Solitons and Compactons). The fundamental property of a soliton is thus to retain its identity in nonlinear interactions. In low-dimensional systems (understood here as environments admitting solitons in one spatial dimension and line solitons in two dimensions), a soliton's amplitude, shape, and velocity are restored after each interaction, whereas in more complex systems this request embraces all dynamically significant qualities. In other words, a structure is a soliton if and only if its *interactions* are *fully elastic*. Thus, the interaction of solitons is always an intrinsic topic whenever solitons are considered.

In the nonlinear world, *linear superposition* does not hold and *nonlinear interaction* takes place. Mathematically, classes of entities are called linear, whenever any linear combination of entities belongs to the same class. Among structures corresponding to the solutions of certain equations, exclusively those that satisfy linear equations are denoted as linear ones. Any linear combination of solutions of a linear equation also solves this equation. The principle of linear superposition is that the resultant structure is simply the sum of the parties and it implies that the individual linear wave shapes are perfectly restored after meeting with each other.

The property of elasticity of soliton interactions has a fundamentally different nature, because, as a rule, a linear superposition of soliton solutions is not a solution of the governing nonlinear equation. This property distinguishes nonlinear *solitary* traveling waves (that is, spatially or temporally localized pulses, whereas the relevant single-wave solution of the underlying equation may be completely stable) from *solitons*, which may be observed and analyzed as separate entities but which always express property (iv) should similar entities appear in the system.

The propagation and interaction of both solitons and solitary waves in realistic conditions (that is, in the presence of dissipation, external fields, and inhomogeneities of the medium) frequently results in a certain loss of energy (usually in the form of radiation of relatively short waves); in certain cases collapse or fission of the structure, or a net exchange of energy between the waves occurs. Even if a solitary wave travels without radiation losses in an ideal environment, it may only be called a soliton if neither radiation loss nor net energy exchange occurs when it meets a similar structure; otherwise a perfect re-emerging of each soliton after the interaction would be impossible. More generally, no net changes of any dynamically or topologically significant quantities are permitted in elastic collisions. Only changes of certain dynamically less significant properties such as the exact location (phase) of the solitons are common in elastic interaction. Solitary waves or structures possessing a selection of properties of solitons are sometimes called quasi-solitons.

Although the first evidence of solitons was extracted from observations in nature, the theory of solitons and their interactions has been developed very much in terms of the mathematical theory of the relevant nonlinear partial differential equations (PDEs) of the motion. The definitions of solitons and their interactions are commonly formulated in terms of solutions to such equations. It is even customary to speak of solitary waves and solitons, and of their interactions, having in mind the relevant exact solutions to these equations. Therefore, in mathematical terms, a soliton – either a solitary wave or a more complex entity – is a solution of a nonlinear PDE that elastically interacts with other similar solutions. In the contemporary nomenclature it is also customary to associate solitons and their interactions with numerical (multi-soliton) solutions to these equations.

Interactions between solitons are the most fascinating features of soliton phenomena. The most instructive quality of soliton interactions is the universality of many of their features that exists in spite of the extremely diverse physical origins of the counterparts. The presentation of the basic conceptual ideas involving soliton interactions, a discussion of several particular cases of such interactions and an overview of observations of the relevant phenomena in natural conditions are mostly limited to effects occurring in the elastic interactions of one-dimensional and line solitons, in which the solitons behave very much like ideal mechanical bodies and which serve as the kernel of the classical concept of solitons and their interactions. Some spectacular features of nearly elastic interactions in higher dimensions are interpreted in terms of generalizations of the classical elastic interactions. Finally, a selection of practical applications of specific features of soliton interactions is again discussed from the viewpoint of line solitons.

Introduction: Key Equations, Milestones, and Methods

Integrable Equations

The quality of some nonlinear PDEs to admit soliton solutions is associated with the property of integrability. Integrable equations usually admit an infinite number of integrals (constants) of motion. Many nonintegrable equations possess localized shape-preserving traveling wave solutions that resemble solitons. However, only integrable equations have the universal property of possessing exact multi-soliton solutions that reflect perfectly elastic interactions between individual solitons. Thus, the integrability of the underlying equation is a primary constituent of the classical soliton interactions. Arnold [10] defines, for example, a soliton as a solitary wave (solution) of an integrable PDE having additional stability and robustness features which are inherited directly from the integrability of this PDE. Nonlinear interactions even between the greatly different (but physically relevant) solutions of integrable equations are also elastic.

Among the variety of integrable PDEs that admit multi-soliton solutions, examples related to the Korteweg–de Vries (KdV) equation, the (focusing) nonlinear Schrödinger (NLS) equation, the integrable Boussinesq equation, the sine-Gordon (SG), and the Kadomtsev–Petviashvili (KP) equation will be used below. Below we shall refer to the listed equations as the classical soliton equations. Since the KdV and the KP equations represent a large number of different physical systems and their solutions can be easily visualized in terms of the easily accessible environment of shallow-water waves, solutions to and results from studies of these equations will be largely used for illustrations of the explanations below.

The rapid development of the theory of solitons has led to the discovery of many integrable equations which show multi-soliton solutions and describe soliton interactions (see, e. g. Arnold [10] and the entry ► Partial Differential Equations that Lead to Solitons). A number of analytical methods have been developed to obtain the (multi-) soliton solutions of integrable equations, starting from the late 1960s. A major tool is the *Inverse Scattering Transform* (IST), first developed by Gardner, Green, Kruskal and Miura [47,50] for the KdV equation. The method consists of associating to the evolution equation a Sturm–Liouville problem. In the case of the KdV equation the Sturm–Liouville equation is just the time independent Schrödinger equation of quantum mechanics, where the potential is the wave function of the KdV equation. The single- and multi-soliton solutions are associated with the discrete spectrum of eigenvalues. A generalization of the Inverse Scattering Transform, known as the *AKNS method* (named after Ablowitz, Kaup, Newell and Segur [2] who developed the method), applies to a large class of integral equations. See the entry ► Inverse Scattering Transform and the Theory of Solitons for detailed information.

At the same time, Hirota [62] found that soliton solutions can be obtained by reducing, through a suitable transformation, the evolution equation to a *bilinear form*. Using some properties of the bilinear operator, it is possible to find the multi-soliton solution of an integrable equation. The proof of integrability of a particular PDE is usually nontrivial. The experience with the Hirota method, applied to a variety of nonlinear wave equations, is that exact multi-soliton solutions are found in many cases when it is not known if the equation is integrable. It is natural to expect that the existence of such solutions reflects the integrability in some sense; however, no proof of this conjecture is known at this time.

Mathematically speaking, it is always possible to build an integrable equation representing certain features of an observed phenomenon. Nevertheless, only a few of these equations can be properly derived, normally using asymptotic expansions, from the corresponding primitive equations describing physical phenomena. Even the classical soliton equations only approximately describe the natural phenomena. A more exact description of the natural processes requires the introduction of additional contributions to these equations. Such contributions usually destroy the property of integrability and lead to nonelasticity of interactions of soliton-like entities. Consequently, most of the times solitons are just a good approximation of solitary waves in nature. This limitation is to some extent balanced by the fact that in many cases integrable soliton-admitting equations adequately represent the basic features of physical phenomena even far beyond their formal scope of validity. In spite of such intrinsic limitations, the tools developed in soliton theory have allowed researchers to reach a very deep understanding of some physical phenomena which would have hardly been explained by other means.

Milestones

Several milestones of the solitons' history (▶ Solitons: Historical and Physical Introduction) are particularly significant to soliton interactions. Already the famous first observations in 1834 of shallow-water solitons [132] implicitly involved certain effects created through soliton interactions. The observed wave ahead of a ship probably had a straight crest, as the one reproduced in 1995 in the Union Canal near Edinburgh (see, e. g. p. 11 in [34]). It is a remarkable feature of ship-induced solitons in relatively narrow and shallow channels that a straight-crested upstream soliton exists ahead of a two-dimensional (2D) disturbance, whereas the solitons generated ahead of a ship sailing in wider shallow water areas have curved crests. The straight crest of the Russell's soliton and of her sister structures are thus not intrinsically one-dimensional (1D) structured solitons. They are formed owing to a specific mechanism of soliton interactions as described below. The KP equation, allowing a simple but adequate description of the crest-straightening phenomenon, was derived even later than the word soliton was coined.

A substantial step towards the concept of solitons was made in the mid-1950s by Fermi, Pasta and Ulam [40] who numerically analyzed the evolution of phonons in an anharmonic lattice. From the mathematical point of view this problem can be described by a discretization of the KdV equation. Surprisingly at the time, the process did not lead to an energy equipartition among the modes, although nonlinearity generally tends to create such an equipartition. Instead, a sort of recurrence, an early evidence of the elastic soliton interactions, was observed.

The basic features of soliton interactions were first demonstrated for media described by the KdV equation [64,174]. The evolution of a sinusoidal initial wave with periodic boundary conditions gives rise to an ensemble of solitary waves that move with different speeds and gradually overtake each other. Originally eight of them were mentioned; later revisitations increased the count to nine [106]. The difference in the number of solitons is not principal since small-amplitude solitons only become evident through additional phase shifts in extremely long-term simulatons [39,133]. One surprise of the system was that after a very long time the whole profile reappears. The other, conceptually much more striking effect, was the persistence of the waves: after a certain phase of interaction with each other, they continued thereafter as if there had been no interaction at all. This persistence, which reflects the essence of the soliton interactions, led Zabusky and Kruskal to invent the name "soliton".

Solitons and new features of their interactions were later discovered in a large number of different physical systems. The universal principle generalizing the physical forces affecting the birth of all solitons is that the propagating disturbance (a pulse in a fiber, a wave-packet on a water surface, a self-focusing optical beam, a localized vortex, or a vortex dipole, etc.) is captured in a potential well jointly induced by its motion and by virtue of the nonlinearity of the underlying system. The solitons can thus be interpreted as the bound states of the relevant potential wells, and soliton interactions – as interactions between the bound states of a jointly induced potential well, or between bound states of different wells located at close proximity. This concept of potential well becomes vividly evident in the inverse scattering theory due to the conservation of the eigenvalues of the underlying spectral problem. This universality explains why, in spite of the immense diversity of the solitons and the underlying physical systems, the interactions (collisions) between solitons in all of these systems follow the same principles.

Classical Soliton-Admitting Equations and Appearance of Solitons

The particular appearance of soliton interactions may vary a lot, depending on the physical system and the nature of the solitons. The Russell soliton is an example of nontopological solitons that often become evident in the form of traveling waves and that cannot exist in rest. On the contrary, single topological solitons (that interpolate between two different states of the system which have the same energy) may exist in rest.

A localized entity consists of an infinite number of Fourier harmonics and generally experiences (linear) spreading or wave radiation. A localized entity of permanent form therefore may only appear if the spreading effects are balanced by a certain restoring force that evidently can only be of a nonlinear nature. In other words, all solitons correspond to a certain balance between spreading and nonlinear effects, the latter becoming evident, for example, as nonlinear steepening of the wave shape in the KdV system, or focusing of diffractive structure in nonlinear optics.

The physical meaning of spreading depends on the particular system. Dispersion is responsible for the spreading of pulses in all media in which the group velocity depends on the wave properties. The classical examples of dispersive solitons are surface [132], internal [33] and Rossby solitons [128] in geophysical systems. Similar structures occur in a large variety of environments such as drift solitons or ion-acoustic solitons in plasma [169], or solitons in optical fibers [60].

Many examples of dispersive solitons and their interactions occur in the framework of the KdV equation and its generalizations. It is a characteristic equation governing weakly nonlinear, spatially one-dimensional (1D) waves whose phase speed attains a simple maximum for infinitely long waves

$$\tilde{\eta}_t + \frac{3c_0}{2h}\tilde{\eta}\tilde{\eta}_x + \frac{c_0 h^2}{6}\tilde{\eta}_{xxx} = 0\,,$$

$$\text{in nondimensional variables}\quad \eta_t + 6\eta\eta_x + \eta_{xxx} = 0\,, \tag{1}$$

where η is the relative water surface elevation in the shallow water environment (Fig. 1), h is the still water depth, and c_0 is the maximum phase speed (celerity) of linear waves at this depth. Its generalization to the case of solitons propagating in slightly different directions

$$(\eta_t + 6\eta\eta_x + \eta_{xxx})_x + 3\eta_{yy} = 0\,, \tag{2}$$

was derived by Kadomtsev and Petviashvili [66]. Named the KP equation after them, it describes features of a so-called weakly 2D environment (also called 1.5-dimensional systems). The nondimensional (x, y, t, η) and physical variables $(\tilde{x},\ \tilde{y},\ \tilde{t},\ \tilde{\eta})$ in Eq. (2) are related as follows: $x = \sqrt{\varepsilon}(\tilde{x} - \tilde{t}\sqrt{gh})\big/h$, $y = \varepsilon\tilde{y}\big/h$, $t = \sqrt{\varepsilon^3 gh}\,\tilde{t}/h$, $\eta = 3\tilde{\eta}\big/(2\varepsilon h) + O(\varepsilon)$ whereas $\varepsilon = |\tilde{\eta}_{\max}|\big/h \ll 1$.

Since the KdV equation only admits nontopological elevation solitons, the KdV/KP framework misses several basic features of soliton interactions, such as collisions of solitons with different topological charge and interactions of solitons of different type. These classes of dispersive solitons are represented by the sine-Gordon (SG) equation

$$\theta_{tt} - c_0^2\theta_{xx} + \omega_0^2 \sin\theta = 0\,, \tag{3}$$

where θ in many applications has the meaning of a certain 2π-periodic angle and ω_0 is the minimum angular frequency of linear waves. This equation, with its name stemming from a play of words regarding its form which is similar to the Klein–Gordon equation, arose already in the 19th century in studies of certain surfaces. It grew greatly in importance when it was realized that it led to kink and antikink solutions with the collisional properties of solitons [116]. Its major field of application is solid state physics, yet it also appears in a number of other physical environments such as the propagation of fluxons in Josephson junctions (a junction between two superconductors), the motion of rigid pendulums attached to a stretched wire, and dislocations in crystals. Its kink solutions (Fig. 1)

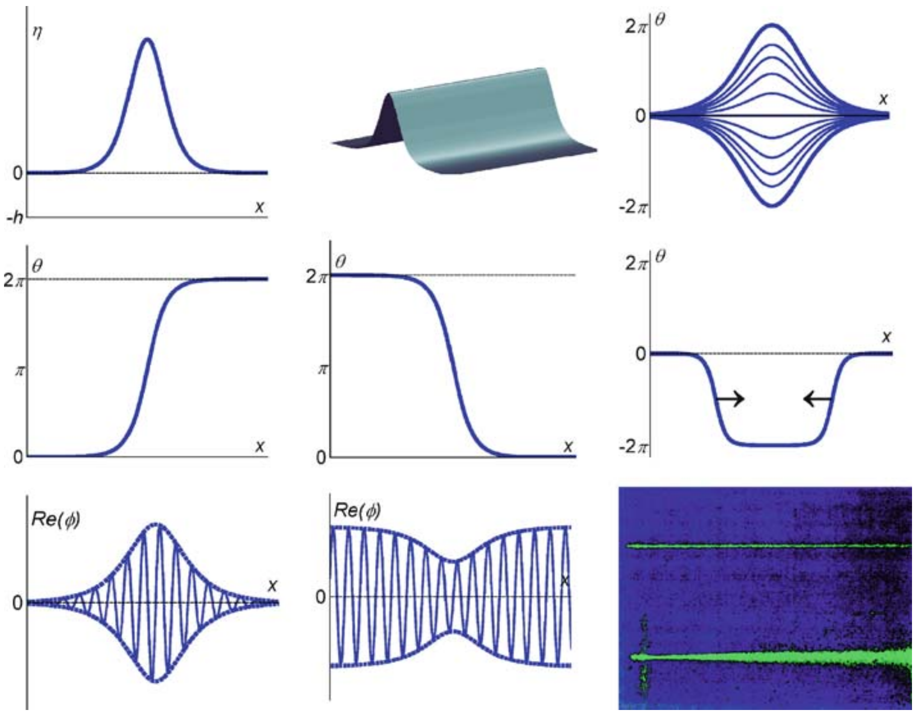
$$\theta_S = 4\arctan\exp\left(\pm\frac{\omega_0}{c_0}\frac{x - vt - x_0}{\sqrt{1 - v^2/c_0^2}}\right)\,, \tag{4}$$

where x_0 is the initial position of the single soliton at the x-axis and $v < c_0$ is the speed of translation of the soliton, represent so-called topological solitons that interpolate between two different states of the system, which have the same energy and may exist in rest (e. g. a frozen (anti)kink, $v = 0$). The SG equation is a simple model admitting breather solitons and soliton pairs with different topological charges. The bulk topological charge is an invariant of the system. In a more complex manner the topological charge becomes evident in the theory of optical spatial solitons, where it can be interpreted in terms of spin of elementary particles. The kink and antikink solutions also represent the possibility of the existence of solitons and antisolitons. Both attractive and repulsive interactions can be vividly demonstrated in this framework and, after all, interactions of physically meaningful solitons of different types are possible.

Another key equation for dispersive waves is the nonlinear Schrödinger (NLS) equation that has a nondimensional form

$$i\psi_t + p\psi_{xx} + q|\psi|\psi^2 = 0 \tag{5}$$

and only has soliton solutions in the focusing case $pq > 0$. This is a universal equation describing the evolution of complex wave envelopes ψ in a dispersive weakly nonlinear medium. It applies, among other phenomena, to



Solitons Interactions, Figure 1

A selection of solitons. From *left* to *right*: *top* – KdV traveling wave soliton, KP plane (*line*) soliton, SG breather; *middle* – kink, antikink, and a kink-antikink pair; *bottom* – bright and gray envelope soliton, and a top view photograph of a 10-μm-wide optical spatial soliton propagating in a strontium barium niobate photorefractive crystal and the same beam diffracting naturally. The *bottom right* image is from [152]. Optical spatial solitons and their interactions: Universality and diversity. Reprinted with permission from AAAS

deep water waves. The studies of its soliton solutions and their interactions also date back to the 1960s [175,176]. Its major application in the context of soliton interactions is nonlinear optics, where its different versions describe the propagation and interactions of both dispersive solitons (e. g. envelope solitons in optical fibers) and diffractive optical spatial solitons.

The basic reason for the existence of optical solitons is that the optical properties (refractive index or absorption) of some materials are modified by the presence of light. This feature introduces nonlinearity into the system and alters the propagation of optical pulses either in space or in time. A dispersive optical soliton is formed when the linear dispersion effects are balanced by a sort of lensing (of short light pulses of permanent form that are called temporal solitons in the optical nomenclature) in an appropriate material (for example, an optical fiber). They were predicted by Hasegawa and Tappert [60] and first observed by Mollenauer et al. [97]. The basic properties of their interactions [59] are analogous to those occurring for KdV or SG solitons. There is continuous interest for their studies because of their applications, e. g. in long-distance optical communication systems.

The generic example of diffractive solitons, evolution and interactions of a part of which are described by the NLS equation in two space dimensions, are optical spatial solitons. The natural tendency to broaden in space owing to the diffraction of optical beams (Fig. 1) can be balanced, for example, due to the optical Kerr effect that consists of a local refractive index change induced by light. As a first approximation, this change is assumed to take place instantaneously and to be proportional to the local intensity of light. In the focusing case (when an increase of the intensity of light causes an increase in the refractive index), the light-induced lensing can make an optical beam stable with respect to perturbations in both its width and intensity. The resulting beam in a 2D or 3D medium is an example of a diffractive soliton. Interaction of such beams occurs owing to overlapping of the modified regions of the medium. Usually such an interaction is local and has a long range only under specific conditions [129].

The first studies suggesting the existence of such solitons in nonlinear optical Kerr media date back to the 1960s [11,27]. In some cases (such as the 1D Kerr solitons in a planar medium) they are classical solitons and described by a fully integrable equation (the NLS equation in two space dimensions [176]). Self-trapped light beams and more complex structures in a 2D and 3D medium attracted the attention of researchers starting from the 1980s, when the appropriate materials were found [13].

In multi-dimensional media, both diffraction and dispersion affect the propagation of solitary waves. A classical example of a diffractive-dispersive system is the motion of short-crested waves on the water surface. Most of the analysis of soliton interactions in this system has been made for effectively 1D solitons with infinitely long crests (so-called *plane* or *line solitons*); the effects of diffraction being implicitly neglected in their propagation and interactions. A practically realized example of diffractive-dispersive solitons are short pulses of incoherent (or white) light that is self-trapped both in the direction of its propagation and in the transversal direction(s), and that propagate changing neither their shape nor their length [94].

Extended Definitions

The classical nomenclature of low-dimensional solitons and their interactions was formulated several decades ago, and it is not surprising that many later discoveries have led to attempts to extend its content. A part of the extensions add several consistent features (such as the conservation of linear and angular momentum, and topological charge or spin during the soliton interactions); however, in several schools substantial generalizations of the term soliton and of the interpretation of the soliton interactions have been introduced. The only component of the classical nomenclature kept in all the interpretations is the essential role of nonlinearity. Since it is virtually impossible to reflect all the extensions, mostly the material and results following the classical definition are presented below.

The solitons are commonly interpreted as a subset of traveling solitary waves. The classical nomenclature favors structures that are stationary in a proper coordinate system; yet it does not exclude the resting topological solitons (e. g. single-kink solutions to the sine-Gordon equation).

A principal extension of the classical definition of solitons that caused quite a serious discussion in the 1970s [101] consists of accounting for the *resonant interactions*. They may result in the fusion of several solitons into a new soliton or, equivalently, in the decay of solitons into counterparts [177]. Although the number of solitons is not conserved, all the parties of the resonant system are solitons at the limit of exact resonance and no energy radiation occurs. It is commonly accepted now as a variation (or a limiting case) of elastic interactions.

The classical soliton interactions preserve the shape of the solitons. The shape is understood in a wide sense; e. g. the shape of the envelope of the NLS solitons is preserved as well as the time-dependent shape of breathers that have an internal oscillation. This quality, which is universal for all scalar solitons, has a more general interpretation in the case of *vector* (composite) *solitons*. For example, the otherwise elastic interactions of Manakov vector solitons exhibit a shape-changing nature [125] in an integrable system consisting of two coupled NLS equations [82]. This property has a paramount practical importance and bright perspectives e. g. in optical soliton-based computations ([155], see also entry ▶ Optical Computing), being one of the key options of practical applications of specific features of soliton interactions.

An obvious reason for a wider view on soliton interactions is that the relatively simple integrable equations admitting soliton solutions stem from certain asymptotic expansions and only approximately describe the processes in nature. A more exact representation of the physical processes through inclusion of higher order terms of those expansions generally destroys the integrability of these equations. In many cases, though, the perturbing terms are small. The influence of a small dissipative perturbation is usually obvious: it brakes the moving solitons and/or damps the oscillating solitons. Such structures are sometimes called dissipative solitons although such a name is somewhat self-contradicting. Effects due to conservative (Hamiltonian) perturbations are much richer in content. Usually they do not destroy or brake solitons, but they may, e. g., render otherwise elastic soliton interactions to inelastic ones owing to extra wave radiation. A well-known example of the effect of perturbations is dissipation-induced annihilation of a kink-antikink pair in a perturbed sine-Gordon equation that otherwise would lead to a breather [114]. A monumental survey of the relevant problems is presented in [77]; see also entry ▶ Soliton Perturbation.

Since none of the physical and numerical environments perfectly matches the governing equations, solitary waves both in realistic conditions and in numerical simulations always are approximations to solitons. In numerical experiments practically ideal solitons can be represented and errors basically occur due to purely computational inaccuracies. Laboratory experiments with solitons encounter problems with perfect excitation of even single solitons and moreover with controlling their interaction properties; however the relevant results are of fundamental importance in establishing the adequacy, the limits and the practical applicability of the properties of solitons and their interactions.

In the theory of elementary particles frequently a soliton is defined merely as a localized solution (resp. entity) of permanent form and the constraint of its survival in collisions is released (e. g. Rebbi [127]). A popular and deep example is the quantum field theory, where e. g. the Yang–Mills field equations admit solutions localized in space (which represent very heavy elementary particles), and also solutions, localized in time as well as in space (instantons). This position has certain support by the conjecture that the classical soliton equations can be interpreted as reductions of self-dual Yang–Mills equations and thus the classical solitons form a subset of the above solutions.

A similar position is also widely adopted in studies of optical beams and vortices. A large part of such beams are described by nonintegrable equations [135] and thus lie beyond the classical soliton nomenclature. Many features of their behavior and interactions, however, demonstrate a striking similarity with that of classical solitons, and exhibit universal properties [152], resembling the similar universality of the classical solitons interactions. Soliton-like beams organized by different nonlinear effects were mostly called solitary waves until the mid-1990s. The modern nonlinear optics nomenclature treats all self-trapped optical beams as solitons [136] although their interactions are not always elastic. A selection of results of studies of their interactions that give a flavor of the richness of soliton interaction phenomena in 3D media is presented below.

The constraint of energy conservation both in the motion and the interactions of the solitons is frequently disregarded in studies of long-lived structures which exhibit reasonable radiation [42] but that still survive in certain collisions. On the other hand, other highly interesting solitary structures such as (Rossby) dipole modons [83,84] do not radiate energy but only survive under certain conditions. A certain amount of results and concepts in this domain is presented to highlight the overlapping of the properties of soliton interactions and interactions of other long-living structures.

Elastic Interactions of One-Dimensional and Line Solitons

The term soliton was originally introduced for nonlinear disturbances, the interaction of which resembles the collision of particles and is fully elastic (Table 1), so that the number of solitons is always conserved and their amplitudes are fully restored afterwards. In low-dimensional cases such as those described by the KdV equation, the amplitudes, directions and propagation velocities of each soliton always recover their initial values (Fig. 2). The shape of each soliton is prescribed by the structure of (single-) soliton solutions to this equation and, as the simplest evidence of the recurrence of the system, it is not surprising that the shape is perfectly restored as well.

Three phases of the interactions of solitons can be distinguished either in time or in space (Fig. 2). First, well-defined, clearly separated solitons approach each other. In the second (interaction) phase, they usually lose their identity and merge into a composite structure. After a while, the solitons emerge again. The appearance of the composite structure depends on the particular physical system; as a rule it is neither a linear superposition of (properly shifted) counterparts nor a new soliton. Therefore solitons survive the collision event, even though they completely merge for a while. The variety of the composites range from the complete vanishing of the counterparts (e. g. kink-antikink collision in the SG equation) over certain interactions of KdV solitons, in which the individual identities are almost conserved and two humps are always visible, to a fourfold elevation of the water surface in a resonant interaction of the KP solitons.

In some environments, one cannot say whether the solitons propagate through each other as waves do or collide as particles do. Both interpretations have a solid ground. The interaction of unidirectional KdV solitons resembles the collision of two particles whereas, e. g. in head-on interactions of Rossby dipole solitons[84] the identity of both counterparts can be continuously tracked and the interaction process resembles a sort of complex overtaking.

Attraction and Repulsion

Another feature of soliton interactions resembling collisions of real massive and/or electrically charged particles that exert certain forces on each other is that soliton interactions may show an *attractive* or *repulsive* nature. This feature can be demonstrated in the framework of the kink-antikink solution of the sine-Gordon equation

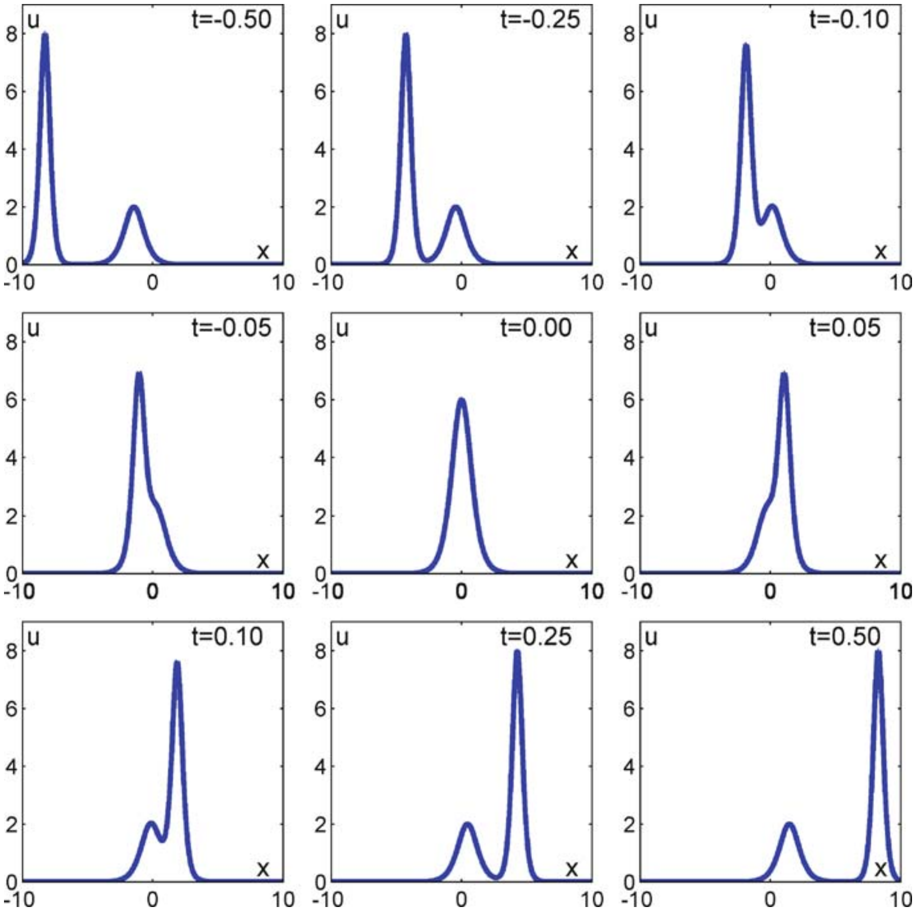
$$\theta_{SA} = 4 \arctan \frac{c_0 \sinh v t \xi}{v \cosh x \xi}, \quad \text{where} \quad \xi = \frac{\omega_0}{c_0} \frac{1}{\sqrt{1 - v^2/c_0^2}} \tag{6}$$

and that only exists if $v > 0$. The motion of the counterparts speeds up when they move towards each other, and slows down when they move apart after passing each other. If one examines the positions of the centers of the counterparts of this solution, the sequences of their positions in the vicinity of their meeting point are not the same

Solitons Interactions, Table 1

Properties of solitons in elastic interactions of 1D and line solitons. Analogous rules hold for angular momentum and spin in elastic collisions in higher dimensions

Property	1D interactions and nonresonant interactions of line solitons	Exceptions in resonant interactions
Number of solitons	Conserved except in a short phase of interaction	Changed
Energy	Restored after interactions for each soliton	Merging possible; total energy conserved.
Amplitude Shape (steepness)	Substantial local changes may occur in the interaction region; restored after interactions	Durable changes may occur
Phase or location	Finite phase shifts commonly occur; yet no phase shifts in certain environments. The total phase shift is exactly equal to the sum of partial shifts that would result from separate collisions with each of the solitons	Infinite phase shifts
Geometry	Substantial changes in the interaction region, restored after interaction	Durable changes may occur
Propagation direction, linear momentum	Conserved for each soliton	Conservation of bulk linear momentum
Topological charge	Conserved for each soliton	



Solitons Interactions, Figure 2

Temporal evolution of KdV solitons described by the two-soliton solution $u(x,t) = 12 \times (3 + 4\cosh(2x - 8t) + \cosh(4x - 64t))/([3\cosh(x - 28t) + \cosh(3x - 36t)]^2)$ [37]. The amplitudes of the counterparts are $u_{1\,\max} = 8$ and $u_{2\,\max} = 2$. Notice the gradual decrease of the taller soliton when it catches the shorter, and its gradual increase when it moves further ahead of the shorter one, and the relatively small amplitude $\tilde{u} = 6$ of the composite structure

as they would be if they had been alone in the system, or far from the other counterpart.

If the solitons were massive or charged particles, one would say that an attractive force exists between them which decreases as their separation increases. On the contrary, the speed of two kinks as well as two antikinks approaching each other slows down as their separation decreases, which is characteristic for repulsive interaction. Another classical reflection of "forces" between solitons is the effect of attraction or repulsion of (envelope) solitons in optical fibers [53].

In two dimensions, the attractive interactions of two obliquely interacting line solitons become evident as an attraction of the relevant wave crests. In the case of optical spatial solitons the attraction affects their centroids. For repulsive interactions the opposite holds. The relevant effects are vividly demonstrated in the framework of the KP equation for the description of drift vortex solitons in environment, otherwise described by the Hasegawa–Mima equation [158]. This environment is equivalent to the one described by the Charney–Obukhov equation for large-scale motions in geophysical hydrodynamics admitting Rossby waves and solitons.

Transient Amplitude Changes, Durable Phase Shifts and Recurrence Patterns

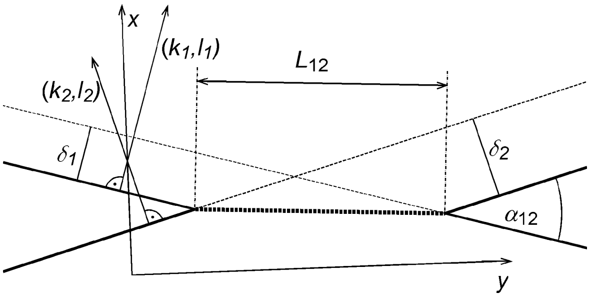
Already the first study of solitons [174] revealed some other universal aspects of soliton interactions. First, certain durable phase shifts may occur, the magnitude of which depends on the interaction details. For example, during the collision presented in Fig. 2, the taller soliton is shifted forward by an amount $\Delta x_1 = \frac{1}{2}\log 3$ and the shorter one backwards by $\Delta x_2 = \log 3$ compared to the case when the solitons are alone in the system. The above-discussed "forces", if present in the system, may be interpreted as the cause of the phase shifts.

The total phase shift of a soliton induced by elastic collisions with any number of solitons is exactly equal to the sum of (partial) shifts that would result from separate collisions with each of the solitons involved. This property is commonly referred to as the *absence of many-particle effects*.

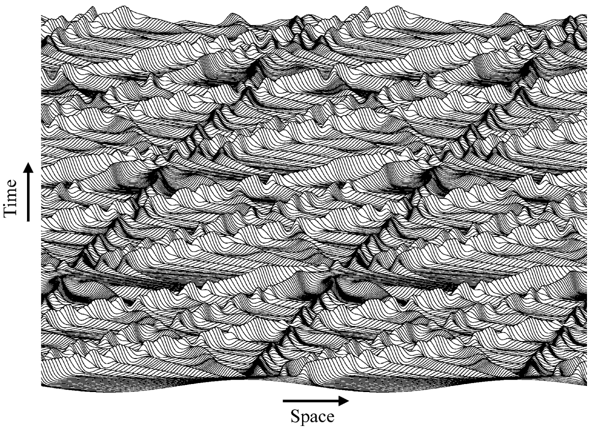
In the case of line solitons the phase shifts become evident in the form of durable shifts of the counterparts during the interaction (Fig. 3). In the negative phase shift case the solitons are delayed compared to their position in the absence of interactions, whereas in the positive phase shift case they seem to be accelerated.

Second, the amplitudes of the interacting solitons may gradually change as they approach to each other (Fig. 2). The amplitude of the composite structure is generally not equal to the sum of amplitudes of the interacting solitons. In interactions of a small number of KdV solitons running along an infinite medium, the behavior of each counterpart and its influence on the partners can always be identified from the shape of the elevation since after some time the solitons become separated enough to detect all of them and their perfectly restored properties in the limit of infinite separation.

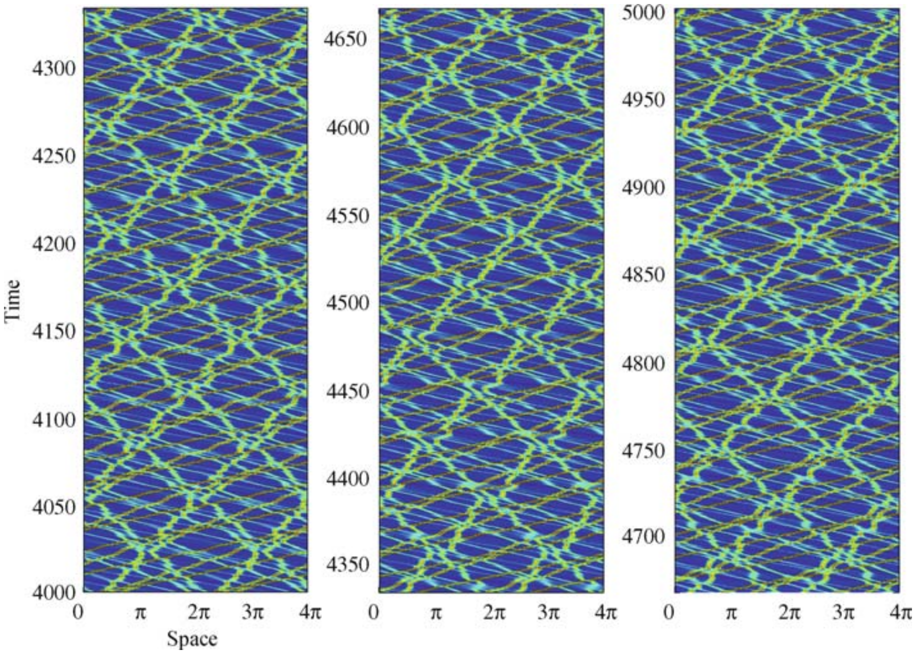
More insight into the properties of the soliton interaction is provided by long-term simulations of ensembles of solitons excited from a periodical signal (Fig. 4). The total number of interacting parties is infinite then, the solitons are tightly packed and such a separation not necessarily occurs. In some cases the presence of several shorter solitons can only be identified through their contributions



Solitons Interactions, Figure 3
Idealized patterns of crests of interacting KP solitons (*bold lines*), their position in the absence of interaction (*dashed lines*) and the crest of the residue s_{12} (*bold dashed line*, see Eq. (7)) corresponding to the negative phase shift case (adapted from [120])



Solitons Interactions, Figure 4
Time-slice plot over two 2π periods of the evolution of the ensemble of the KdV solitons generated from a sinusoidal initial condition in space for $0 \le t \le 97$ at $\log d = -2.3$. Reprinted from [134]



Solitons Interactions, Figure 5

Patterns of locations of crests of the KdV solitons generated from sinusoidal initial conditions at log d = −2.3209. Reused with permission from [39]

to the phase shifts of the partners since they are seldom visually identifiable; such solitons are called hidden solitons [39].

The trajectories of the (crests of the) counterparts may show quite a complex pattern and may exhibit both simple recurrence (at which the initial system is approximately restored) as well as super-recurrence (understood as a nearly perfect restoration of the initial system over longer times).

Practically periodical in time rhombus-like patterns (suggesting that interactions of quite a small number of solitons govern the properties of the system) optionally occur in cases when a super-recurrence is possible (Fig. 5), whereas in other cases these patterns show gradual variation also over extremely long times. There seems to exist a critical value d^* in the range $-1.8 < \log \mathrm{d}^* < -1.9$ of the dispersion parameter d in the KdV equation (presented in the form $\eta_t + \eta\eta_x + \mathrm{d}\eta_{xxx} = 0$), which distinguishes if a super-recurrence exist. If $\mathrm{d} < \mathrm{d}^*$, no super-recurrence can be detected even within extremely long calculation times [39,134].

Durable Local Amplitude Changes in Oblique Interactions of Line Solitons

The interaction of unidirectional KdV solitons does not create any large changes in the wave amplitudes ([37], see also ▶ Korteweg–de Vries Equation (KdV), History, Exact *N*-Soliton Solutions and Further Properties of the). However, amplitude amplification may occur under certain conditions when line solitons propagating in slightly different directions meet each other. Extensive evidence comes from coastal engineering, where it was known already a long time ago that ocean waves approaching breakwaters and seawalls under a certain angle caused unexpectedly large overtopping. This phenomenon was systematically studied already in the 1950s [117], well before the word soliton was coined.

When a shallow-water wave is obliquely launched along an ideally reflecting wall, the reflection does not necessarily follow the rules of geometrical optics. For a certain range of incidence angles, the crests of the approaching and the reflected wave fuse together near the wall and the process resembles the *Mach reflection*. The common crest, an analogue of the *Mach stem*, is generally unsteady and gradually lengthens for sine waves [15] and Stokes waves [173]. This type of reflection also occurs during perfect reflection of random [89] and solitary waves of different nature [45,92,93,115,159].

For a certain set of parameters of the approaching KdV soliton, the resulting structure is equivalent to half of the pattern created by two interacting KP solitons of equal amplitude (Fig. 3). The prominent feature of both the (Mach) reflection and oblique interactions of shallow-water solitons is that the height of the common crest may considerably exceed the sum of the heights of the interacting solitons. For interacting waves with equal amplitudes the hump can be up to four times as high as the incoming waves. Originally established in the context of the reflection of Boussinesq solitons ([92,93]; named Mach reflection by Melville 1980 [91]), the amplitude amplification in a certain region occurs in all oblique attractive interactions of line KP solitons (that is, in interactions with a negative phase shift). Since the interaction pattern is stationary in an appropriately moving coordinate system for a range of the parameters of the interacting solitons, it may lead to durable local changes of the amplitude of the resulting pattern of elevation.

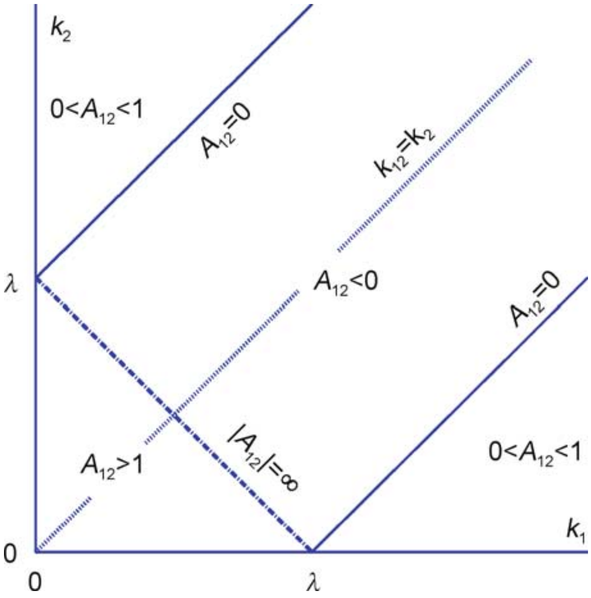
Resonance

While the theory of integrable systems requires that the 1D solitons retain their identity in interactions, collisions of solitons in multiple dimensions may lead to the formation of new solitons. The phenomenon called *resonant interaction* leads to the emergence of new structures that combine the energy of the counterparts and to modifications of the number of solitons and of some of their geometrical features.

This possibility was first recognized by Newell and Redekopp [101] in the context of the KP and NLS equations. Miles [93] gives an early answer to their question about the practical consequences of this phenomenon in terms of an essential increase of the wave amplitude occurring during the resonant Mach reflection. Originally named "breakdown of the Zakharov–Shabat theory of integrable systems with more than one spatial dimension", it highlights the more complex nature of soliton interactions in more than one dimension, where large phase shifts are possible.

Remarkably, the resonance conditions $\boldsymbol{\kappa}_\infty = \boldsymbol{\kappa}_1 + \boldsymbol{\kappa}_2$, $\omega_\infty = \omega_1 + \omega_2$, where $\boldsymbol{\kappa}_i = (k_i, l_i)$, $i = 1, 2, \infty$ are the wave vectors and ω_i are the frequencies of the solutions of the corresponding linearized KP equation, are precisely the same as those for the resonant interaction of three weakly nonlinear waves. Conceptually, however, such interactions more resemble so-called double resonance, in which not only the above resonance conditions are met but also the group velocities of two or more counterparts are equal and otherwise sporadic energy exchange within the resonant triplets is replaced by much more intense interactions (e. g. [145]).

The patterns of line soliton crests (Fig. 3) suggest that only attractive interactions may result in resonance. The resonant interactions between line solitons correspond to infinitely large phase shifts and infinitely strong attraction, and lead to durable changes of the geometry of the system. In the situation depicted in Fig. 3, the resonance would lead to an infinitely long common crest and thus render the "four-arm" pattern to a "three-arm" one.



Solitons Interactions, Figure 6
Negative and positive phase shift areas for the two-soliton solution of the KP equation for fixed $l_{1,2}$. The phase shift $\Delta_{12} = -\ln A_{12}$ has a discontinuity along the line $k_1 + k_2 = \lambda$ at which $|A_{12}| = \infty$ and the resonant interactions occur. No two-soliton solution exists in the area where $A_{12} < 0$. An analogue of linear superposition occurs when $A_{12} = 0$. Adapted from [120]

In the KP framework, the resonance phenomenon is connected with the fact that its multi-soliton solutions representing the interaction of line solutions only exist for a subset of the space of parameters of the interacting solitons (Fig. 6). Resonance occurs when the parameters of the interacting solitons lie at a certain part of the border of this subset.

A new soliton born at the exact resonance may resonantly interact with yet another soliton. This mechanism has been employed in [17] to construct a family of exact solutions to the KP equation corresponding to resonance of several solitons. Such solutions consist of unequal numbers of incoming and outgoing line solitons, the parameters of which can be obtained from asymptotic analysis [16]. This class contains a variety of multisoliton solutions, many of which exhibit nontrivial spatial interaction patterns that are generally characteristic to multisoliton solutions to the KP equation [118].

Geometry of Oblique Interactions of KP Line Solitons

Patterns

A relatively simple but instructive, easily visualizable and rich-in-content model demonstrating the discussed features of interactions of line solitons is the KP equation. It admits simple explicit formulae for multi-soliton solutions and extensive analytical treatment of their geometrical properties. The two-soliton solution to the KP equation can be written as a sum $\eta = s_1 + s_2 + s_{12}$ of terms, reflecting to some extent the two interacting solitons s_1, s_2 and residue s_{12}

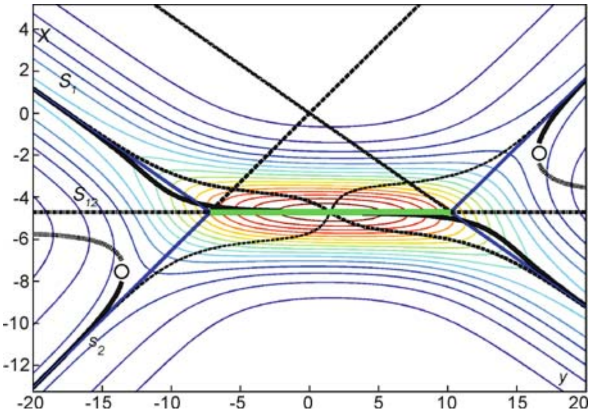
$$\begin{aligned} s_{1,2} &= \sqrt{A_{12}} k_{1,2}^2 \Theta^{-2} \cosh\left(\varphi_{2,1} x + \ln\sqrt{A_{12}}\right) , \\ s_{12} &= 2\Theta^{-2}\left[(k_1 - k_2)^2 + A_{12}(k_1 + k_2)^2\right] , \\ \Theta &= \cosh\frac{\varphi_1 - \varphi_2}{2} + \cosh\frac{\varphi_1 + \varphi_2 + \ln A_{12}}{2} . \end{aligned} \tag{7}$$

Here $\varphi_i = k_i x + l_i y + \omega_i t,\ a_i = \frac{1}{2}k_i^2,\ i = 1, 2,$ are the phases and amplitudes of the interacting solitons, the 'frequencies' ω_i satisfy the dispersion relation $k_i\omega_i + k_i^4 + 3l_i^2 = 0$ of the linearized KP equation, $A_{12} = [\lambda^2 - (k_1 - k_2)^2]/[\lambda^2 - (k_1 + k_2)^2]$ is the phase shift parameter and $\lambda = l_1 k_1^{-1} - l_2 k_2^{-1}$.

The two-soliton solution only exists in a part of the parameter space $R^{(4)}(k_1, k_2, l_1, l_2)$. An interaction may result in either the positive or the negative phase shifts $\delta_{1,2} = -\ln A_{12}/|\boldsymbol{\kappa}_{1,2}|$ of the counterparts (Fig. 6). The negative phase shift (cf. Fig. 3) can be attributed to an attractive interaction and the positive phase shift to a repulsive interaction.

The residue s_{12} (with an amplitude a_{12}) is at times considered as a virtual 2D "interaction soliton" [118], which grows into the resonant soliton at the limit of exact resonance $A_{12} \to \infty$ (cf. Fig. 7). This virtual structure can be heuristically interpreted as supported by both dispersive and diffractive effects, the latter being balanced by the oblique motion of the interacting solitons.

The patterns of both idealized (Fig. 3) and factual (Fig. 7) wave crests are symmetric with respect to a particular point called interaction center, and are stationary in a properly moving coordinate frame. If the amplitudes of the solitons are equal, the pattern is equivalent to the reflection of the incoming soliton from the wall following the motion of the interaction center. For unequal amplitude solitons the equivalence is not complete because mass and energy flux occur through such a wall.



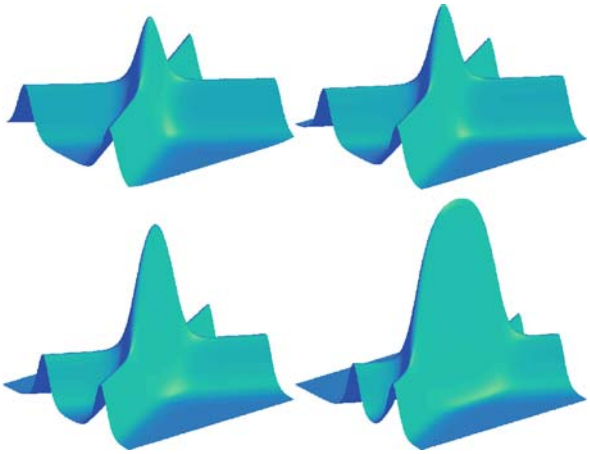
Solitons Interactions, Figure 7
Patterns of idealized (*solid/dashed straight lines*) and factual (*dashed curves*) crests of the incoming solitons s_1, s_2, the crest of the residue s_{12} (*green line*), the visible wave crests and troughs (*bold* and *bold dashed curves*, respectively), and isolines of the two-soliton solution in the case $l = l_1 = -l_2 = 0.3$, $k_1 = 0.6507$, $k_2 = 0.4599$ in normalized coordinates (x, y). The interaction center is the crossing point of all factual crests. Contour interval is $\frac{1}{4}\max(a_1, a_2) = 0.0689$. *Circles* show the common points of the troughs and crests of the two-soliton solution. Reprinted from [146]

Amplitudes

The phase shifts $\delta_{1,2}$ of the counterparts only depend on the amplitudes of the incoming solitons and the angle $\alpha_{12} = 2\arctan(\frac{1}{2}\lambda)$ between their crests. For the negative phase shift case $A_{12} > 1$ (that is typical in interactions of solitons with comparable amplitudes, Fig. 5), an interaction pattern emerges, the height of which exceeds the sum of the heights of two incoming solitons. The maximum surface elevation for equal amplitude solitons is $a_{\max} = 4a_{1,2}/(1 + A_{12}^{-1/2})$; thus, the nonlinear superposition of solitons may lead to a fourfold amplification of the surface elevation. In a highly idealized case of interactions of five solitons the maximum surface elevation may exceed the amplitude of the incoming solitons by more than an order. The largest elevation occurs if the interacting solitons are in exact resonance $A_{12} = \infty$, when solitons intersect under a physical angle $\tilde{\alpha}_{12} = 2\arctan\sqrt{3\tilde{\eta}/h}$. For unequal amplitude solitons the maximum elevation $a_{\max}$ for finite A_{12} and the amplitude of the resonant soliton a_∞ at $A_{12} = \infty$ are [46,146]

$$a_{\max} = a_{12} + 2A_{12}^{1/2}\frac{a_1 + a_2}{\left(A_{12}^{1/2} + 1\right)^2}, \quad a_\infty = \frac{(k_1 + k_2)^2}{2} . \tag{8}$$

The near-resonant high hump is particularly narrow and its front is very steep. The maximum slope of the front of the two-soliton solution may be eight times as large as



Solitons Interactions, Figure 8
Surface elevation in the vicinity of the interaction area, corresponding to the incoming solitons with equal amplitudes intersecting under different angles [120]

the slope of the incoming solitons [150]. For unequal amplitude solitons, the amplification of slope of the front of the interaction pattern is roughly proportional to the amplitude amplification. The extraordinary steepness of the front of the near-resonant hump, although intriguing, is not unexpected, because the resonant soliton is higher and therefore narrower than the incoming solitons.

The length L_{12} of the idealized common crest (Fig. 3) is $L_{12} \sim \ln A_{12}$ and therefore is modest unless the interacting solitons are near-resonant. The length of the area, where the total elevation exceeds the sum of the amplitudes of the counterparts, may be considerably longer; however it is also roughly proportional to $\ln A_{12}$ [150]. For largely different amplitudes of the interacting solitons, the amplitude amplification remains modest but the spatial extent of substantial influence of nonlinear interaction is roughly as large as if the amplitudes were equal [151]. Their near-resonant interaction becomes evident as a wave with its crest oriented and propagating differently compared with the counterparts. The described features are universal for that part of the wave vector space, in which the two-soliton solution with a negative phase shift exists. In the case of a positive phase shift (that is typical for interactions of solitons with largely different amplitudes) the highest elevation does not exceed the amplitude of the larger soliton.

Numerical simulations of the process of formation of the composite structure show that a high wave hump is formed quite fast in the interaction region and its height soon considerably exceeds the sum of the heights of the counterparts [58,124]. A high and gradually lengthening wave hump is also formed in cases when no exact two-soliton solution of the KP equations exists. The interaction in such cases is inelastic: neither the orientation of the crests nor the height of the solitons before and after interaction match each other, and a certain amount of wave radiation takes place. The described features become evident in a number of simulations of soliton interactions in different environments [105,168].

Soliton Interactions in Laboratory and Nature

Ion-Acoustic Solitons

The most productive studies of soliton interactions in the 1970s and the 1980s were performed in the framework of ion-acoustic waves. Their behavior is approximately described by the KP equation. An analytical solution describing interactions and resonance of two plane ion-acoustic solitons of small amplitude in the 3D collisionless plasma was given in [171]. This solution is wider than the two-soliton solution of the KP equation in the sense that the nearly unidirectional approximation is not used.

The first successful experiment proving that the 2D interaction of solitons has a wider spectrum of features compared with the 1D collisions is reported in [178]. A collision of two equal amplitude spherical solitons led to significant changes and to the emergence of a new 2D object. This interaction was not perfectly elastic: the solitons were not only shifted in phase but also reduced in amplitude. Soon afterwards, a near-resonant interaction of two KP solitons, with a large-amplitude wave hump of the counterparts oriented across the principal propagation direction, was observed near the interaction center, whereas the counterparts experienced clear phase shifts [43]. The amplitude amplification was found to be close to twofold the sum of the amplitudes of the incoming solitons [103].

The first adequate evidence of the nonlinear increase in amplitude during collisions of unequal amplitude KP solitons was also obtained in this framework [46]. For largely different amplitudes the measured maximum amplitudes were less than the predicted ones. A probable reason for the discrepancy is that the near-resonant wave hump occurs and covers a substantial area (equivalently, can be reliably estimated with the use of relatively large probes) only if the incoming solitons are very close to resonance. The common part of the crests of interacting solitons in the reported experiments was pretty short suggesting that the solitons were not exactly in resonance. Moreover, the complete identification of all the crests in the interactions of solitons with considerably different amplitudes is not possible because the crests are partially invisible (Fig. 7, [146]). The generic reason for experimental difficulties in this and similar experiments is the above-discussed feature that large-amplitude structures tend to mask the presence of shorter ones.

Shallow-Water and Internal Solitons

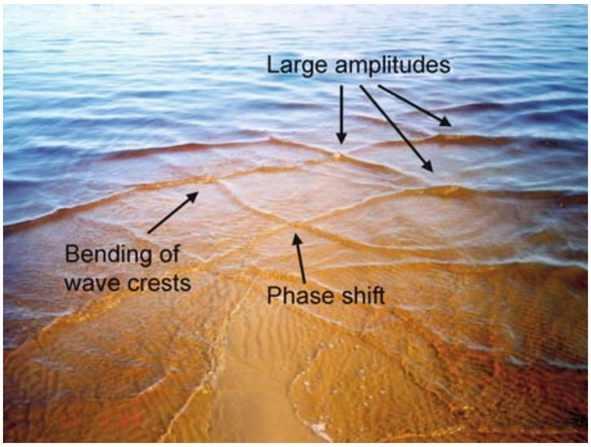
Phase shifts, formation of a common crest of the interacting soliton-like waves, and accompanying amplitude and orientation changes are commonly observable in very shallow areas (Fig. 9). A famous photo by Terry Toedtemeier (1978), in which two solitonic shallow-water waves have significant phase shifts and quite a long common crest on a beach in the US state of Oregon, has been reproduced in a number of sources (e. g. p. 24 in [34]).

Although [93] predicted up to fourfold amplification of the surface elevation, much weaker amplitude amplification was found in early experiments [91]. An amplification close to the theoretical one was detected as a by-product of studies of shallow-water channel superconductivity [25].

Russell's "great wave of translation" [132] probably had a practically straight crest, reflecting a well-known feature of the generation of solitons in relatively narrow channels: a nearly perfectly 1D upstream soliton is created also by a 2D disturbance. The KdV model, which is frequently used to describe this phenomenon, is intrinsically 1D and certainly results in straight wave crests. An analysis of the simplified problem of ship motion in the KP equation and the Boussinesq equation shows, in accordance with common experience, that the wave crests are curved [71]. In fact, the essentially 2D waves with paraboloidal crests generally emerge in front of the disturbance. Their curvature monotonously decreases as the waves propagate [85,87].

Although the sidewall of the channel is not essential for the radiation of the upstream solitons, still it plays a crucial role in the transformation of already radiated waves. The straight-crested solitons are formed in the process of the (Mach) reflection of solitons with curved crests from the sidewalls [113]. This feature can be distinguished in many numerical studies, where an initially curved soliton starts to straighten when it comes in touch with the sidewall [149]. Since the reflection of solitons, which in this case is equivalent to soliton interactions, plays the key important role in creating straight-crested waves in channels, with a certain exaggeration it can be said that Russell's soliton exists due to soliton interactions.

Probably the most common solitonic phenomenon in nature next to the shallow water waves is represented by internal solitary waves in the oceans and in the atmosphere ([56,107,111,130], see also the entry ► Non-linear Internal Waves). In many cases, the KdV equation is an appropriate model for internal waves and the effects occurring during their interactions are equivalent to those described above.



Solitons Interactions, Figure 9
Interaction patterns of solitonic surface waves in very shallow water near Kauksi resort on Lake Peipsi, Estonia. Photo and copyright by Lauri Ilison, July 2003. First published in [151]

Large-scale internal solitary waves and their groups exhibit many features unique to soliton interactions and are often called simply internal solitons. Intersecting solitary wave packets on Synthetic Aperture Radar (SAR) images frequently exhibit phase shifts while crossing each other [51], similarly to phase shifts occurring in the KP model, where the agreement between theory and experiments is quite good. Apel et al. [9] summarize advances in the descriptions and observations of internal solitons and their interactions.

The other environment for internal solitons, which provides some instructive aspects and is eligible, e. g., for the description of small-scale internal solitons in a two-layer medium, is the Benjamin–Ono (BO) equation [1]

$$\eta_t + c\eta_x + \alpha\eta\eta_x + \frac{\beta}{\pi}\frac{\partial^2}{\partial x^2} \quad P.V. \left[\int_{-\infty}^{\infty} \frac{\eta\left(x', t\right)}{x - x'} \mathrm{d}x' \right] = 0. \tag{9}$$

First, the BO solitons do not acquire a phase shift after a collision. Second, this environment demonstrates that generalization of integrable 1D soliton-admitting equations to even a weakly 2D case may lead to the loss of integrability, thus the preservation of this feature in the KP equation is not universal. The weakly 2D analogues of the BO equation [1] apparently are not integrable [9] and do not admit simple analytic multi-soliton solutions. Numerically simulated oblique interactions of weakly 2D BO solitons show that a phenomenon resembling the resonant interaction of the KP solitons and an accompanied amplitude amplification do occur but the newly generated wave in the zone of nonlinear interaction is far from the BO soliton [105,167].

Solitons at Planetary Scale

Large-scale internal waves and solitons in nature are affected by the joint influence of the Earth's rotation and sphericity, and perfect internal line solitons generally cannot exist in the ocean. The physical reason of this "antisoliton theorem" is that due to rotational dispersion, there is always a phase synchronism between a source moving at an arbitrary speed and linear perturbations in such environments. This leads to unavoidable wave radiation from a large-scale solitary internal wave [49]; however this effect is negligible for relatively small-scale solitons, the typical time scales of which in the oceans are a few tens of minutes.

The existence of nonstationary solitary waves is still possible [52]. They exhibit a phenomenon similar to the recurrence of solitons – the repeating decay and reemergence process, and formation of a nearly localized wave packet consisting of a long-wave envelope and shorter, faster solitary-like waves that propagate through the envelope [61]. The robust, long-lived structure may contain as much as 50% of the energy in the initial solitary wave. Interacting packets may either pass through one another, or merge to form a longer packet. Their further studies may reveal interesting features concerning the functioning of oceans and other large stratified water bodies. Yet in the majority of practically interesting cases such long-term behavior is masked by topographic effects that modify the internal waves much more dramatically.

Another medium of particular practical importance as well as an instructive one from the viewpoint of soliton interactions is the large-scale geophysical flow in the oceans and the atmosphere. It is frequently treated as anisotropic quasi-2D turbulence, which has the property of energy concentration in large-scale structures. Eddies in such flows are stabilized by the background rotation and also affected by the joint effect of the Earth's rotation and sphericity. The latter becomes evident through the North-South variation of the Coriolis force, the so-called (planetary) beta-effect. In this environment, large-scale vortices often maintain their coherence for surprisingly long times and show an amazing variety of interactions; perhaps the most well-known example being the Jovian Great Red Spot.

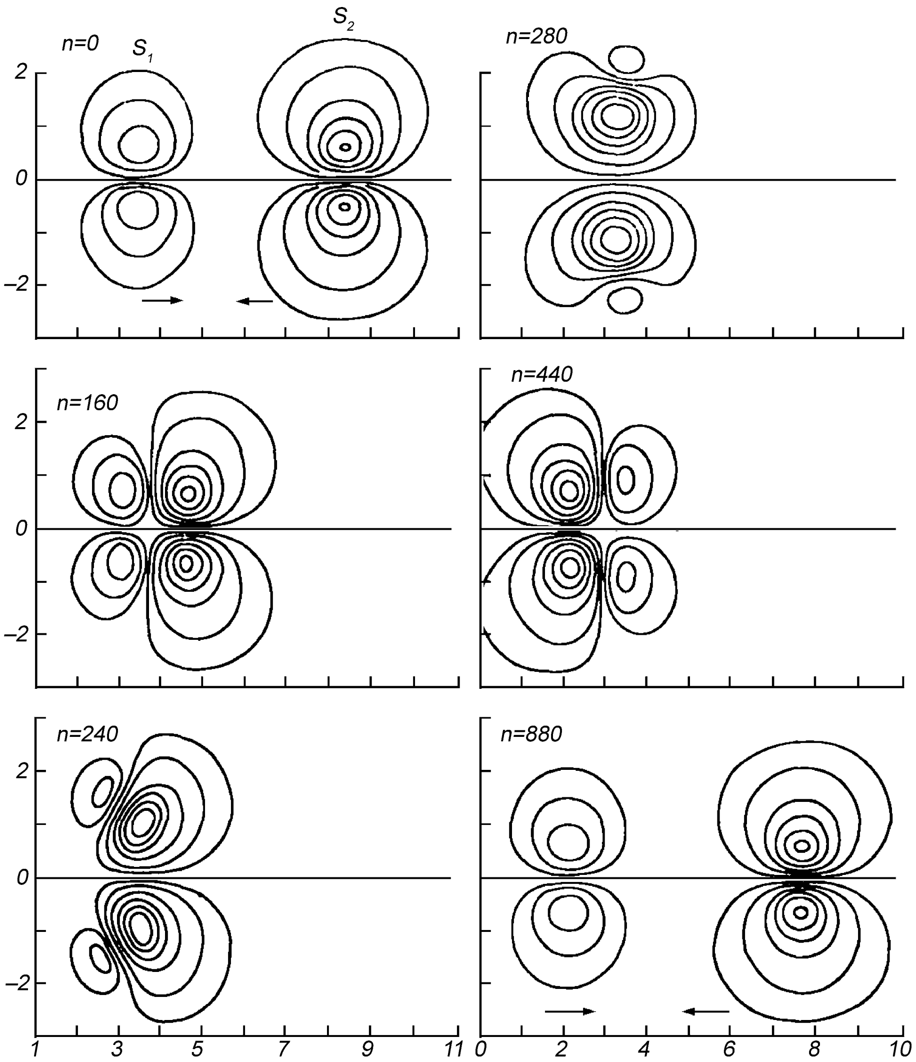
Form-preserving, uniformly translating, horizontally localized solutions (of the equations of planetary dynamics) are called modons. In nondissipative quasi-geostrophic dynamics they exhibit a variety of properties of solitons. They carry a certain amount of fluid with them, which distinguishes them from (solitary) waves that cause, if at all, a finite displacement of the medium. The word modon was coined a decade after soliton was first used [157].

The standard quasi-geostrophic models predict that the accompanying Rossby-wave radiation should cause a decay of isolated eddies. Several models explain the persistence of monopolar vortices using external features of the flow, or interactions with a background shear that suppresses the Rossby-wave radiation [19,126,144]. A number of generalizations of the quasi-geostrophic model predict that, within a certain range of parameters, a nonlinear anticyclonic vortex of a special form can exist for a long time [121,153,170] due to the mutual compensation of weak (scalar) nonlinearity and weak dispersion. Some authors call such structures (Petviashvili-type) Rossby solitons, although their collisions are not elastic.

Long-living coherent vortices and their interactions have been studied in many experiments with rapidly rotating parabolic vessels [102]. The increase of the effective gravity due to the centrifugal acceleration towards the side walls modifies the governing equations, so that the (Petviashvili-type) Rossby soliton does not exist in the paraboloidal geometry [104]; yet anticyclonic vortices have a very long lifetime [154]. Another option to study the evolution of such vortices is a tank with a sloping bottom to imitate the planetary beta-effect (e. g. [41,90], among others; for an overview of earlier studies see [63]).

The majority of interactions of solitary vortices in such an environment are inelastic. The 2D localized vortices of comparable size and intensity, and like sign attract each other and generally coalesce soon. Only practically equivalent vortices may form a quasi-stable pair rotating around each other. The vortices of different signs either repulse from each other or form a relatively stable dipole [54,55]. The simultaneous evolution of a large number of soliton-like vortices has mostly been studied theoretically or numerically (e. g. in terms of a cluster of point vortices or hetons, [57]).

The simplest stable barotropic modon in flat-bottom beta-plane dynamics is the Larichev–Reznik dipole [83]. It is a traveling dipolar vortex with a characteristic north/south antisymmetry. It propagates eastward at any speed, or westward at speeds greater than the longest-wave speed. The speed of translation is therefore out of the range of the phase speed of Rossby waves, which is the reason why this modon does not radiate waves (unlike large-scale internal solitons). The significance of such structures in everyday



Solitons Interactions, Figure 10
Head-on collision of Larichev–Reznik dipoles [67]

life and climate dynamics (where they may be responsible, e. g., for certain unusual weather events since the vortex pair may behave completely differently compared with a single cyclone or anticyclone) is far from being well understood.

A part of the numerically simulated direct head-on and overtaking collisions of Larichev–Reznik modons are elastic [84]. An instructive feature of the head-on collisions is that the identity of all four vortices can be sometimes tracked during all the interaction phase (Fig. 10). The process resembles a collision of soft particle pairs connected by an elastic chain. Their oblique collisions generally are not elastic and may lead to the formation of new dipoles, tripoles, or to the destruction of the structures.

Effects in Higher Dimensions

Optical Spatial Solitons

Starting from the mid-1980s, various optical solitary entities have become a key arena for studies of the properties and interactions of solitary structures in multiple dimensions. The interest in these studies is continuously supported by a variety of existing and emerging applications of such solitons and their interactions in communication and computation technology (see also the entry ► Optical Computing). Another reason for the rapid progress in these studies is the relative ease of their generation in suitable materials, combined with the precision of contemporary optical experiments and the possibilities of precise control over almost every parameter [136,152]. They offered a lot of the further conceptual progress of soliton interactions through making possible 3D propagation of solitary entities in laboratory conditions, albeit in many cases these entities did not match the classical nomenclature of solitons. They were called solitary waves in the field for several decades, because the governing equations are not necessarily integrable and the interactions not necessarily elastic. Still their interaction at times shows amazingly elastic nature and they are commonly called optical spatial solitons (OSS) in the modern nonlinear optics nomenclature [136].

A large number of various kinds of OSSs have been identified [76,135]. Observations of spatial Kerr solitons became possible from the mid-1980s, when so-called slab waveguides were built [3,13]. Their behavior substantially depends on the dimensions of the soliton and the waveguide. For example, the bright Kerr solitons are stable only in planar systems whereas the 2D solitons were found to collapse. Quasi-stable 2D solitons in Kerr media described by the cubic NLS in two spatial dimensions exist in the form of so-called necklace-ring beams [143].

Optical spatial solitons occurring due to the saturation of the nonlinear change in the refractive index were first demonstrated in the 1970s [18] and realized in a solid medium in the 1990s [74]. The net effect of the multiple physical effects involved in photorefractive materials is that the underlying nonlinearities are also saturable. Such solitons were also discovered in the early 1990s [137] and were found to be stable in both slab waveguides [75] and in bulk media [38].

So-called quadratic solitons (that consist of multifrequency waves coupled by means of a second-order nonlinearity) were predicted in the mid-1970s [70] and realized experimentally in the 1990s [164]. They can be thought of as vector solitons because they involve mutual self-trapping of two or more components [152]. Their interactions are similar to those occurring in other saturable nonlinear media.

Coherent and Incoherent Collisions

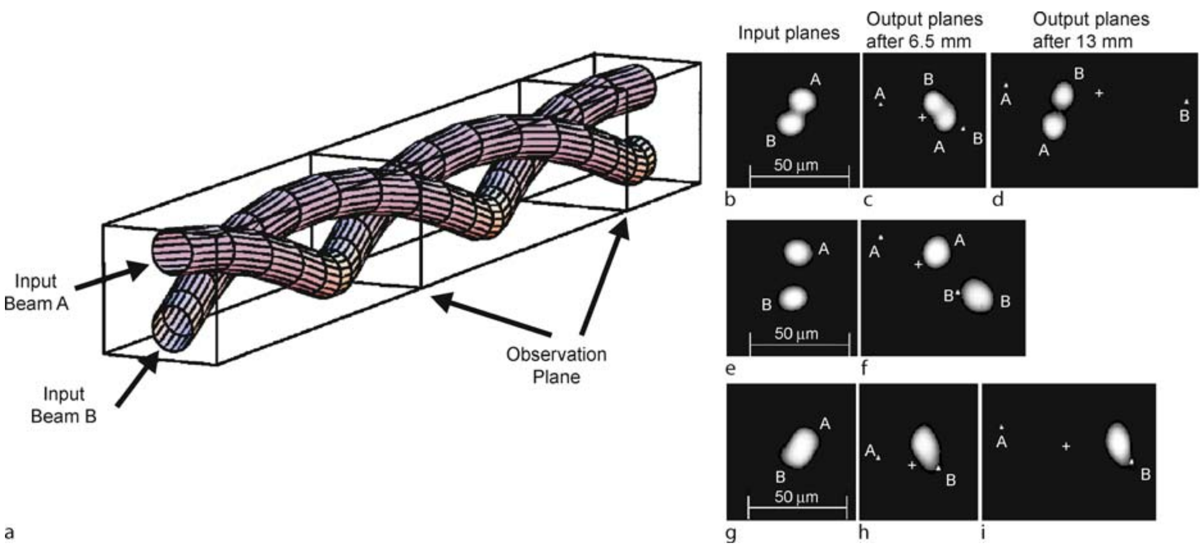
Coherent interactions of OSSs occur when the nonlinear medium responds quickly to the (interference) effects in the overlapping area of the beams. The increase of the intensity of light in the overlapping region of two parallel launched equivalent in-phase solitons in the focusing medium leads to an increase in the refractive index in that region. This in turn forces the centroid of each soliton toward it. Hence the solitons attract each other. The spatial distribution of the intensity of light resembles the temporal distribution of the amplitude of attracting interacting line solitons. In the case of antiphase beams, an opposite process occurs and such solitons exert repulsion. Coherent collisions of Kerr solitons have been demonstrated and the attraction for in-phase and the repulsion for antiphase solitons were clearly observed at the beginning of the 1990s [4,5,139]. The collision of nonequivalent coherent OSSs results in an energy exchange between the beams and in a repulsive force that makes the beams diverge.

Some reactions (e. g. photorefractive and thermal) of the optical medium are relatively slow. Solitons within which the phase varies randomly across the beam were first demonstrated in [95] and later extended to the beams of incoherent white light [94]. So-called incoherent soliton interactions occur when the response of the medium only follows the average intensity of light. Interactions of incoherent bright solitons are always attractive (Andersen and Lisak 1995).

The long-distance behavior of the counterparts in attractive interactions depends on the properties of the medium and the orientation of the beams. Pure Kerr solitons launched along nearly parallel directions result in periodic paths of the solitons' centroids. As in the case of the recurrence of the KdV solitons, Kerr solitons exactly return to their initial conditions after each cycle. Solitons launched under large enough converging angles exhibit a slight lateral deflection (which is equivalent to the phase shift of the KdV solitons). The solitons launched under large enough diverging angles never collide.

If the energy transfer were neglected, the force between the equivalent 1D Kerr solitons would vary from the maximum attractive between in-phase solitons to the maximum repulsive for antiphase solitons. This process leads to the (generally nonperiodic) energy transfer, where one soliton may gather net energy from the other. The details of the trajectories and other properties of the interacting solitons can be quite complex [152].

Interactions of 2D OSSs (e. g. in media with saturating nonlinearities) have a much larger variety of scenarios [152]; for example, interactions of incoherent photorefractive solitons may be both attractive and repulsive [156]. A number of resonance-like inelastic phenomena have been observed in which the number of solitons is not conserved. The *fusion* of two or more solitons into one structure resembles the resonance phenomenon. Here it happens owing to the gradual change of the orientation of the waveguides of attracting solitons launched under small relative angles into a new waveguide [48]. Differently from the resonance of line solitons, this happens for a certain range of initial parameters. The threshold is the maximum total internal reflection angle in the induced



Solitons Interactions, Figure 11

a The experimentally observed soliton spiraling process. The arrows indicate the initial trajectories. b, e, and g show different input conditions and c, f, and h are the outputs after 6.3 mm and d and i after 13 mm. The triangles indicate the centers of the corresponding diffracting beams. From [152]

waveguide [140,142]. For larger collision angles, the solitons interact as described above. The solitons can fuse together already on the first merging or after a certain number of oscillations of their trajectories with decreasing amplitudes and periods [12,78,160,161]. A sort of inverse of this process is the breakup (*fission*, [142]) of optical solitons into new solitons upon their interaction [163]. *Annihilation* of solitons upon collision also may occur, when three solitons collided and only two emerged from the collision process [79].

Three-Dimensional Effects

In 3D space the trajectories of interacting beam solitons do not necessarily lie on a single plane and the system possesses nonzero initial angular momentum. The trajectories of solitons that are launched along skew lines passing close to each other may bend in three dimensions. The solitons *spiraling* around each other form a double helix orbit [123], much like two celestial objects or two moving charged particles moving along nearly parallel trajectories do. The interaction generally leads to spiraling-fusion or spiraling-repulsion of the trajectories [14,160,161]. Since the mutual rotation encounters a centrifugal force (which is always repulsive), a perfect double-helix DNA-like system, an interesting class of elastic interactions, requires soliton attraction. It may be formed from identical in-phase coherent solitons; yet such a double helix is unstable with respect to perturbations of both the relative phase and amplitude of one of the solitons. This effect resembles processes occurring during the interactions of vortices of different and like sign in quasi-2D rotating fluids.

Attractive interactions of incoherent solitons have made possible the observation of spiraling solitons in saturable media [141]. The two solitons of equal power orbit periodically about each other. Such a pair conserves angular momentum and is more or less stable for a certain range of the initial orientation and distance between the beams. Although the underlying saturable nonlinearities are described by nonintegrable equations, the spiraling process does not lead to measurable energy radiation [20]. When the initial distance between the solitons is increased, they do not form a long-living interacting system. The double spiral for beams located initially very close to each other converges fast and the counterparts eventually fuse. The pair apparently has a limited lifetime. A likely reason for its fusion or off-spiraling is the potential coherence of the light in the beams.

Interactions of Vector Solitons

Vector (composite) solitons consist of two or more components (also called modes) that mutually self-trap in a nonlinear medium. The simplest vector solitons, first suggested by Manakov [88], consist of two orthogonally polarized components in a nonlinear Kerr medium in which self-phase modulation is identical to cross-phase modulation. They are described by an integrable system and interact elastically. An appropriate nonlinear material and experimental conditions for their demonstration were developed a long time after their discovery [68]. Similar solitons may exist if each field component is at a different frequency, and the frequency difference between components is much larger than the inverse of the nonlinearity relaxation time [138,166], or if the field components are incoherent with one another [22,30]. The energy exchange between the components of the interacting vector solitons takes place without radiative losses [7].

Vector solitons can also consist of different modes of their jointly induced waveguide [29,165]. These multimode solitons may possess multiple humps, may contain both bright and dark components [28,166], may have a certain internal dynamics, and were found to be stable (or weakly unstable) in large regions of their parameter space [96,109,110]. Interactions between vector solitons have, additionally to the above-discussed generic properties of solitons interactions (e. g. [86]), also some unique features. The shape transformations of colliding multimode solitons can lead to two different multimode solitons emerging from the collision process [6,69,80]. This "polarization switching" was predicted initially for Manakov solitons [125].

Composite multimode multihump solitons may carry topological charge in one of the vector components [21, 98]. This charge carried by a soliton of finite energy can be interpreted as spin of real particles. Elastic collisions of such 3D solitons should, ideally, conserve not only energy, and linear and angular momentum [122], but also the equivalent of spin. Two colliding composite solitons carrying opposite spins may form a metastable state that later decays into two or three new solitons. If the solitons interact under a certain critical angle, angular momentum is transferred from spin to orbital angular momentum. Finally, the shape transformation of the vortex component occurs at large collision angles, for which scalar solitons of all types simply go through one another unaffected or only have phase shifts [99,100]. The recent progress in optical vortices, vortex solitons, and their interactions is reviewed in [35].

Applications of Line Soliton Interactions

A generic use of soliton interactions in every case where solitary waves or topological solitons may exist, follows from the definition of a soliton. It consists of establishing if the structure is a solitonic one or not from the properties of their interactions.

The practical use of specific features of soliton interactions was started largely in parallel with the first technologies based on the use of solitons (e. g. optical soliton-based communications) more than a decade ago. The most promising is the technology of ▶ Optical Computing (see the relevant entry), which may be partially based on the interactions of vector solitons [155] and which may open completely new horizons in the architecture of computers.

Soliton Interactions on a Water Surface

Several applications based on properties of elastic soliton interactions (or changes of certain fields during such interactions) have been proposed in the framework of water waves. From the purely geometrical properties of oblique interactions of plane solitons one may extract, e. g., the water wave height. The relations for the phase shifts and for the intersection angle can be reduced to the following transcendental equation:

$$\delta_1 \sqrt{2a_1\left(1+\lambda^2/4\right)} = \pm \ln \frac{\delta_2^2\lambda^2 - 2\left(\delta_2-\delta_1\right)^2 a_1^2}{\delta_2^2\lambda^2 - 2\left(\delta_2+\delta_1\right)^2 a_1^2} \,. \quad (10)$$

If the intersection angle α_{12}, the phase shifts $\delta_{1,2}$ and their signs (or the wave propagation directions) are known, equation (10) uniquely defines the heights of the interacting solitons [118,119].

The phenomenon of drastic increase of surface elevation and slope of wave fronts described in Sect. "Geometry of Oblique Interactions of KP Line Solitons" can be attributed to formation of "freak" or "giant" waves in the ocean, the height and steepness of which are considerably larger than expected based on the classical wave statistics [73]. Their sudden appearance is a feature of paramount importance to and a generic source of danger for navigation and in coastal and offshore engineering. Interactions of envelope (Schrödinger) solitons and breather solutions of the NLS [108] in deep water are their potential source, although the model based on envelope soliton interactions underestimates the probability of large wave formation compared with the fully nonlinear model [31]. Interacting solitary wave groups that emerge from a long wave packet can produce freak wave events and may lead to a threefold increase of the wave slope [32].

The problem of extremely large-amplitude internal waves is by no means less important, as they can pose acute (currently neither well understood nor accounted for) danger for submarines, oil and gas platforms, oil risers and pipelines, and to other engineering constructions in relatively deep water. The problem of "freak" internal waves and their generation due to internal soliton interactions has only recently attracted the attention of researchers [81].

There is a clear potential in the use of the advances concerning the geometry of the line soliton interactions in studies of both internal and shallow water solitons. While phase shifts usually have no dynamical significance, unexpectedly large elevations, extreme slopes, or changes in orientation of wave crests owing to the (Mach) reflection or oblique interactions of solitonic waves frequently cause an acute danger. The particularly high (stem) waves may lead to hits by high waves arriving from an unexpected direction. They may break during propagation into shallow water and attack armor blocks from an unprotected side [89] or overtop the breakwaters in unexpected locations [172]. These effects may be particularly pronounced in the case of ship wakes that often approach seawalls or breakwaters from other directions than wind waves do [148]. The possibility of drastic steepening of the front of the near-resonant structure is immaterial in many physical systems but is a crucial component of danger from freak waves [73], since specifically steep and high waves present an acute danger (e. g. [162]). Even if the steepness of the wave front is not large in absolute terms, relatively steep and long, tsunami-like waves tend to become asymmetric and exhibit unexpectedly large runup heights [36].

The amplitude amplification, however, evidently becomes effective relatively seldom, because two or more systems of long-crested solitonic waves must approach a certain area from different directions, and extreme elevations and slopes only occur if the amplitudes of the interacting solitons, the angle between their crests and the water depth are specifically balanced. The long life-time of the resulting wave hump in favorable conditions [73] may drastically increase the probability of the occurrence of abnormally high waves.

Ship-Induced Solitons and Wave Resistance

An increasing source of solitonic waves is the fast ship traffic in relatively shallow areas. The generation of shallow-water solitons is most effective for large ships sailing at speeds roughly equal to the maximum phase speed $\sqrt{gh}$ of surface gravity waves [87]. The low decay rates and exceptional compactness of the solitonic ship wakes has led to a significant impact on the safety of people, property and craft, unusually high hydrodynamic loads in a part of the nearshore, and to a considerable remote impact of the ship traffic in shallow areas [112,147]. Such a wake has probably caused a fatal accident as far as about 10 km from the sailing line already in 1912 [149]. The interactions of ship-induced solitons may lead to dangerous waves in the vicinity of coastal fairways and harbor entrances in otherwise sheltered areas [120].

A intriguing use of soliton interactions, combining effects occurring during an elastic oblique interaction of shallow-water solitons, followed by an annihilation of solitons and forced antisolitons, has been made for reducing the wave resistance in channels [25] and for catamaran design [26]. In a channel, the bow wave reflected from a side wall is used to cancel the stern wave, whereas the bow wave of one hull of the catamaran is used to cancel the stern wave of the other hull. Its practical use is possible for ships sailing at near-critical speeds, the wake of which may consist of a single soliton. The adequate calculation of the phase shift occurring during its Mach reflection (equivalently, during the interaction of two bow solitons) is a key component in achieving a perfect annihilation of properly timed bow solitons and the sterns' waves of depression.

The use of this effect by adjusting the channel width, the water depth and the ship's speed probably is the first intentional use of the features occurring during soliton interactions in the design of a certain technology [23,24]. At the exact design conditions the reflected bow wave completely cancels the stern wave and leads to a sort of channel superconductivity [25]. The same effect may occur between two ships moving in parallel [65].

Finally, it is interesting to note that already J.S. Russell may have been aware of this possibility. He describes efforts of a spirited horse which, pulling a boat in a canal, had drawn the boat up into its own wave leading to a significant reduction in resistance. The boat owner had noted this, and the event led to 'high-speed' service on some canals in the 1820s and 1830s [131]. We shall probably never know if the decrease of wave resistance occurred due to simple crossing of the critical speed or owing to the channel superconductivity; however, even the remote possibility that the practical use of soliton interactions may have been reported even before the first report about solitons, is remarkable.

Future Directions

Although the theory of soliton interactions has been extensively developed during four decades and its basic features for low-dimensional environments have been well understood, it is still in the stage of rapid development. Its first practical applications appeared only a decade ago and there is evidently room for relevant developments in (geo)physical applications. Analysis of long-term evolution of ensembles of solitons is still a challenge and reveals ever new interesting features even in the simplest soliton-admitting systems such as the KdV equation. Although several formal procedures of building explicit multi-soliton solutions to many equations have been known since the 1970s, studies of properties of (optionally resonant) interactions of several line solitons only started at the turn of the millennium. Many fascinating features of elastic interactions of solitons of different kinds in more complex systems will obviously become evident in the nearest future. The results in this direction may largely widen understanding of the integrability of the underlying equations.

Extremely rapid progress may be expected in studies into properties of long-lived soliton-like structures, in particular, of optical spatial solitons. The classification of their nomenclature and interactions is far from being completed. This environment is the key framework for identifying the new features of soliton interactions in higher dimensions, even though interactions of such structures are not necessarily elastic.

Bibliography

1. Ablowitz MJ, Segur H (1981) Solitons and the inverse scattering transform. SIAM, Philadelphia
2. Ablowitz MJ, Kaup DJ, Newell AC, Segur H (1974) The inverse scattering transform – Fourier analysis for nonlinear problems. Stud Appl Math 53:249–315
3. Aitchison JS, Weiner AM, Silberberg Y, Oliver MK, Jackel JL, Leaird DE, Vogel EM, Smith PWE (1990) Observation of spatial optical solitons in a nonlinear glass wave-guide. Opt Lett 15(9):471–473
4. Aitchison JS, Silberberg Y, Weiner AM, Leaird DE, Oliver MK, Jackel JL, Vogel EM, Smith PWE (1991) Spatial optical solitons in planar glass wave-guides. J Opt Soc Am B-Optical Phys 8(6):1290–1297
5. Aitchison JS, Weiner AM, Silberberg Y, Leaird DE, Oliver MK, Jackel JL, Smith PWE (1991) Experimental-observation of spatial soliton-interactions. Opt Lett 16(1):15–17
6. Akhmediev N, Krolikowski W, Snyder AW (1998) Partially coherent solitons of variable shape. Phys Rev Lett 81(21):4632–4635
7. Anastassiou C, Segev M, Steiglitz K, Giordmaine JA, Mitchell M, Shih MF, Lan S, Martin J (1999) Energy-exchange interactions between colliding vector solitons. Phys Rev Lett 83(12):2332–2335
8. Anderson D, Lisak M (1985) Bandwidth limits due to incoherent soliton interaction in optical-fiber communication-systems. Phys Rev A 32(4):2270–2274
9. Apel JR, Ostrovsky LA, Stepanyants YA, Lynch JF (2007) Internal solitons in the ocean. J Acoust Soc Am 121(2):695–722
10. Arnold JM (1998) Varieties of solitons and solitary waves. Opt Quantum Electron 30:631–647
11. Askar'yan GA (1962) Effects of the gradient of strong electromagnetic beam on electrons and atoms. Sov Phys JETP 15(6):1088–1090
12. Baek Y, Schiek R, Stegeman GI, Baumann I, Sohler W (1997) Interactions between one-dimensional quadratic solitons. Opt Lett 22(20):1550–1552
13. Barthelemy A, Maneuf S, Froehly C (1985) Soliton propagation and self-confinement of laser-beams by Kerr optical nonlinearity. Opt Commun 55(3):201–206
14. Belic MR, Stepken A, Kaiser F (1999) Spiraling behavior of photorefractive screening solitons. Phys Rev Lett 82(3):544–547
15. Berger V, Kohlhase S (1976) Mach-reflection as a diffraction problem. In: Proc 25th Int Conf on Coastal Eng. ASCE, New York, pp 796–814
16. Biondini G, Chakravarty S (2006) Soliton solutions of the Kadomtsev–Petviashvili II equation. J Math Phys 47(3):033514
17. Biondini G, Kodama Y (2003) On a family of solutions of the Kadomtsev–Petviashvili equation which also satisfy the Toda lattice hierarchy. J Phys A-Math General 36(42):10519–10536
18. Bjorkholm JE, Ashkin A (1974) cw self-focusing and self-trapping of light in sodium vapor. Phys Rev Lett 32(4):129–132
19. Busse FH (1994) Convection driven zonal flows and vortices in the major planets. Chaos 4:123–134
20. Buryak AV, Kivshar YS, Shih MF, Segev M (1999) Induced coherence and stable soliton spiraling. Phys Rev Lett 82(1):81–84
21. Carmon T, Anastassiou C, Lan S, Kip D, Musslimani ZH, Segev M, Christodoulides D (2000) Observation of two-dimensional multimode solitons. Opt Lett 25(15):1113–1115
22. Chen ZG, Segev M, Coskun TH, Christodoulides DN (1996) Observation of incoherently coupled photorefractive spatial soliton pairs. Opt Lett 21(18):1436–1438
23. Chen X-N, Sharma SD (1994) Nonlinear theory of asymmetric motion of a slender ship in a shallow channel. In: Rood EP (ed) 20th Symposium on Naval Hydrodynamics. US Office on Naval Research, Santa Barbara, pp 386–407
24. Chen X-N, Sharma SD (1997) Zero wave resistance for ships moving in shallow channels at supercritical speeds. J Fluid Mech 335:305–321
25. Chen X-N, Sharma SD, Stuntz N (2003) Zero wave resistance for ships moving in shallow channels at supercritical speeds. Part 2. Improved theory and model experiment. J Fluid Mech 478:111–124
26. Chen X-N, Sharma SD, Stuntz N (2003) Wave reduction by S-catamaran at supercritical speeds. J Ship Res 47:145–154
27. Chiao RY, Garmire E, Townes CH (1964) Self-trapping of optical beams. Phys Rev Lett 13(15):479–482
28. Christodoulides DN (1988) Black and white vector solitons in weakly birefringent optical fibers. Phys Lett A 132(8–9):451–452
29. Christodoulides DN, Joseph RI (1988) Vector solitons in birefringent nonlinear dispersive media. Opt Lett 13(1):53–55
30. Christodoulides DN, Singh SR, Carvalho MI, Segev M (1996) Incoherently coupled soliton pairs in biased photorefractive crystals. Appl Phys Lett 68(13):1763–1765
31. Clamond D, Grue J (2002) Interaction between envelope solitons as a model for freak wave formations. Part I: Long time interaction. Comptes Rendus Mecanique 330(8):575–580
32. Clamond D, Francius M, Grue J, Kharif C (2006) Long time interaction of envelope solitons and freak wave formations. Eur J Mech B-Fluids 25(5):536–553
33. Davis RE, Acrivos A (1967) Solitary internal waves in deep water. J Fluid Mech 29(3):593–608

34. Dauxois T, Peyrard M (2006) Physics of solitons. Cambridge University Press, Cambridge
35. Desyatnikov AS, Kivshar YS, Torner L (2005) Optical vortices and vortex solitons. Prog Opt 47:291–391
36. Didenkulova II, Zahibo N, Kurkin AA, Levin BV, Pelinovsky EN, Soomere T (2006) Runup of nonlinearly deformed waves on a coast. Doklady Earth Sci 411(8):1241–1243
37. Drazin PG, Johnson RS (1989) Solitons: An introduction. In: Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge
38. Duree GC, Shultz JL, Salamo GJ, Segev M, Yariv A, Crosignani B, Diporto P, Sharp EJ, Neurgaonkar RR (1993) Observation of self-trapping of an optical beam due to the photorefractive effect. Phys Rev Lett 71(4):533–536
39. Engelbrecht J, Salupere A (2005) On the problem of periodicity and hidden solitons for the KdV model. Chaos 15:015114
40. Fermi E, Pasta J, Ulam S (1955) Studies of nonlinear problems. I Los Alamos report LA-1940; (1965) In: Segré E (ed) Collected papers of Enrico Fermi. University of Chicago Press, Chicago
41. Firing E, Beardsley RC (1976) The behaviour of a barotropic eddy on a beta-plane. J Phys Oceanogr 6:57–65
42. Flierl GR, Larichev VD, Mcwilliams JC, Reznik GM (1980) The dynamics of baroclinic and barotropic solitary eddies. Dyn Atmos Oceans 5(1):1–41
43. Folkes PA, Ikezi H, Davis R (1980) Two-dimensional interaction of ion-acoustic solitons. Phys Rev Lett 45(11):902–904
44. Fuerst RA, Canva MTG, Baboiu D, Stegeman GI (1997) Properties of type II quadratic solitons excited by imbalanced fundamental waves. Opt Lett 22(23):1748–1750
45. Funakoshi M (1980) Reflection of obliquely incident large-amplitude solitary wave. J Phys Soc Japan 49:2371–2379
46. Gabl EF, Lonngren KE (1984) On the oblique collision of unequal amplitude ion-acoustic solitons in a field-free plasma. Phys Lett A 100:153–155
47. Gardner CS, Greene JM, Kruskal MD, Miura RM (1974) Korteweg–de Vries equations and generalizations: methods for exact solutions. Commun Pure Appl Math 27:97–133
48. Gatz S, Herrmann J (1992) Soliton collision and soliton fusion in dispersive materials with a linear and quadratic intensity depending refraction index change. IEEE J Quantum Electron 28(7):1732–1738
49. Galkin VM, Stepanyants YA (1991) On the existence of stationary solitary waves in a rotating fluid. PMM J Appl Math Mech 55(6):939–943
50. Gardner CS, Green JM, Kruskal MD, Miura RM (1967) Method for solving the Kortweg–de Vries equation. Phys Rev Lett 19:1095–1097
51. Gasparovic RF, Apel JR, Kasischke ES (1988) An overview of the SAR internal wave signature experiment. J Geophys Res-Oceans 93(C10):12304–12316
52. Gilman OA, Grimshaw R, Stepanyants YA (1996) Dynamics of internal solitary waves in a rotating fluid. Dyn Atmos Oceans 23(1–4):403–411
53. Gordon JP (1983) Interaction forces among solitons in optical fibers. Opt Lett 8(11):596–598
54. Griffits RW, Hopfinger EJ (1986) Experiments with baroclinic vortex pairs in a rotating fluid. J Fluid Mech 173:501–518
55. Griffiths RW, Hopfinger EJ (1987) Coalescing of geostrophic vortices. J Fluid Mech 178:73–97
56. Grimshaw R (2001) Internal solitary waves. In: Grimshaw R (ed) Environmental Stratified Flows. Kluwer, Dordrecht, pp 1–30
57. Gryanik VM, Doronina TN, Olbers DJ, Warncke TH (2000) The theory of three-dimensional hetons and vortex-dominated spreading in localized turbulent convection in a fast rotating stratified fluid. J Fluid Mech 423:71–125
58. Haragus-Courcelle M, Pego RL (2000) Spatial wave dynamics of steady oblique wave interactions. Physica D 145:207–232
59. Hasegawa A (1989) Optical Solitons in Fibers. Springer, Berlin
60. Hasegawa A, Tappert F (1973) Transmission of stationary nonlinear optical pulses in dispersive dielectric fiber: II Normal dispersion. Appl Phys Lett 23:171–172
61. Helfrich KR (2007) Decay and return of internal solitary waves with rotation. Phys Fluid 19(2):026601
62. Hirota R (1971) Exact solution of the Korteweg–de Vries equation for multiple collisions of solitons. Phys Rev Lett 27:1192–1194
63. Hopfinger EJ, van Heijst GJF (1993) Vortices in rotating fluids. Annu Rev Fluid Mech 25:241–289
64. Ikezi H, Taylor RJ, Baker DR (1970) Formation and interaction of ion-acoustic solitons. Phys Rev Lett 25(1):11–14
65. Jiankang W, Lee TS, Chu C (2001) Numerical study of wave interaction generated by two ships moving parallely in shallow water. Comput Meth Appl Mech Engrg 190:2099–2110
66. Kadomtsev BB, Petviashvili VI (1970) The stability of solitary waves in weakly dispersive media. Dokl Akad Nauk SSSR 192:532–541
67. Kamenkovich et al. (1987) Synoptic eddies in the ocean. Russian version. Gidrometeoizdat, Leningrad, pp 124
68. Kang JU, Stegeman GI, Aitchison JS, Akhmediev N (1996) Observation of Manako vspatial solitons in AlGaAs planar waveguides. Phys Rev Lett 76(20):3699–3702
69. Kanna T, Lakshmanan M (2003) Exact soliton solutions of coupled nonlinear Schrödinger equations: Shape-changing collisions, logic gates, and partially coherent solitons. Phys Rev E 67(4):046617 Part 2
70. Karamzin YN, Sukhorukov AP (1976) Mutual focusing of high-power light. Sov Phys JETP 41:414–420
71. Katsis C, Akylas TR (1987) On the excitation of long nonlinear water waves by a moving pressure distribution. Part 2. Three-dimensional effects. J Fluid Mech 177:49–65
72. Kaup DJ, Newell AC (1978) Solitons as particles, oscillators, and in slowly changing media: A singular pertubation theory. Proc R Soc London A 361(4):413–446
73. Kharif C, Pelinovsky N (2003) Physical mechanisms of the rogue wave phenomenon. Eur J Mech B Fluids 22:603–634
74. Khitrova G, Gibbs HM, Kawamura Y, Iwamura H, Ikegami T, Sipe JE, Ming L (1993) Spatial solitons in a self-focusing semiconductor gain medium. Phys Rev Lett 70(7):920–923
75. Kip D, Wesner M, Shandarov V, Moretti P (1998) Observation of bright spatial photorefractive solitons in a planar strontium barium niobate waveguide. Opt Lett 23(12):921–923
76. Kivshar YS, Luther-Davies B (1998) Dark optical solitons: physics and applications. Phys Rep – Rev Sect Phys Lett 298(2–3):81–197
77. Kivshar YS, Malomed BA (1989) Dynamics of solitons in nearly integrable systems. Rev Mod Phys 61(4):768–915
78. Krolikowski W, Holmstrom SA (1997) Fusion and birth of spatial solitons upon collision. Opt Lett 22(6):369–377

79. Krolikowski W, Luther-Davies B, Denz C, Tschudi T (1998) Annihilation of photorefractive solitons. Opt Lett 23(2):97–99
80. Krolikowski W, Akhmediev N, Luther-Davies B (1999) Collision-induced shape transformations of partially coherent solitons. Phys Rev E 59(4):4654–4658
81. Kurkin AA, Pelinovsky EN (2004) Freak waves: Facts, theory and modelling. Nizhny Novgorod State Technical University, Nizhny Novgorod (in Russian)
82. Lakhsmanan M, Rajasekhar S (2003) Nonlinear dynamics: integrability, chaos and patterns. Springer, Berlin
83. Larichev VD, Reznik GM (1976) Strongly non-linear, two-dimensional isolated Rossby waves. Oceanology 16:961–967
84. Larichev VD, Reznik GM (1983) Collision of two-dimensional solitary Rossby waves. Oceanology 23(5):725–734
85. Lee SJ, Grimshaw RHJ (1990) Upstream-advancing waves generated by three-dimensional moving disturbances. Phys Fluids A 2:194–201
86. Leo G, Assanto G (1997) Collisional interactions of vectorial spatial solitary waves in type II frequency-doubling media. J Opt Soc Am B-Opt Phys 14(11):3151–3161
87. Li Y, Sclavounos PD (2002) Three-dimensional nonlinear solitary waves in shallow water generated by an advancing disturbance. J Fluid Mech 470:383–410
88. Manakov SV (1974) On the theory of two-dimensional stationary self-focusing of electromagnetic waves. Sov Phys JETP 38:248–253
89. Mase H, Memita T, Yuhi M, Kitano T (2002) Stem waves along vertical wall due to random wave incidence. Coast Eng 44:339–350
90. Masuda A, Marubayashi K, Ishibashi M (1990) A laboratory experiment and numerical simulation of an isolated barotropic eddy in a basin with topographic β. J Fluid Mech 213:641–659
91. Melville WK (1980) On the Mach reflection of solitary waves. J Fluid Mech 98:285–297
92. Miles JW (1977) Obliquely interacting solitary waves. J Fluid Mech 79:157–169
93. Miles JW (1977) Resonantly interacting solitary waves. J Fluid Mech 79:171–179
94. Mitchell M, Segev M (1997) Self-trapping of incoherent white light. Nature 387(6636):880–883
95. Mitchell M, Chen ZG, Shih MF, Segev M (1996) Self-trapping of partially spatially incoherent light. Phys Rev Lett 77(3):490–493
96. Mitchell M, Segev M, Christodoulides DN (1998) Observation of multihump multimode solitons. Phys Rev Lett 80(21):4657–4660
97. Mollenauer LF, Stolen RH, Gordon JP (1980) Experimental-observation of picosecond pulse narrowing and solitons in optical fibers. Phys Rev Lett 45(13):1095–1098
98. Musslimani ZH, Segev M, Christodoulides DN, Soljacic M (2000) Composite multihump vector solitons carrying topological charge. Phys Rev Lett 84(6):1164–1167
99. Musslimani ZH, Soljacic M, Segev M, Christodoulides DN (2001) Interactions between two-dimensional composite vector solitons carrying topological charges. Phys Rev E 63(6):066608
100. Musslimani ZH, Soljacic M, Segev M, Christodoulides DN (2001) Delayed-action interaction and spin-orbit coupling between solitons. Phys Rev Lett 86(5):799–802
101. Newell AC, Redekopp LG (1977) Breakdown of Zakharov–Shabat theory and soliton creation. Phys Rev Lett 38(8):377–380
102. Nezlin VM, Sneshkin EN (1993) Rossby vortices, spiral structures, solitons: Astrophysics and plasma physics in shallow water experiments. Springer, Berlin
103. Nishida Y, Nagasawa T (1980) Oblique collision of plane ion-acoustic solitons. Phys Rev Lett 45(20):1626–1629
104. Nycander J (1993) The difference between monopole vortices in planetary flows and laboratory experiments. J Fluid Mech 254:561–577
105. Oikawa M, Tsuji H (2006) Oblique interactions of weakly nonlinear long waves in dispersive systems. Fluid Dyn Res 38:868–898
106. Osborne AR, Bergamasco L (1986) The solitons of Zabusky and Kruskal revisited: perspective in terms of the periodic spectral transform. Physica D 18:26–46
107. Osborne AR, Burch TL (1980) Internal solitons in the Andaman Sea. Science 208(4443):451–460
108. Osborne AR, Onorato M, Serio M (2000) The nonlinear dynamics of rogue waves and holes in deep-water gravity wave trains. Phys Lett A 275(5–6):386–393
109. Ostrovskaya EA, Kivshar YS, Chen ZG, Segev M (1999) Interaction between vector solitons and solitonic gluons. Opt Lett 24(5):327–329
110. Ostrovskaya EA, Kivshar YS, Skryabin DV, Firth WJ (1999) Stability of multihump optical solitons. Phys Rev Lett 83(2):296–299
111. Ostrovsky LA, Stepanyants YA (1989) Do internal solitons exist in the ocean? Rev Geophys 27:293–310
112. Parnell KE, Kofoed-Hansen H (2001) Wakes from large high-speed ferries in confined coastal waters: Management approaches with examples from New Zealand and Denmark. Coastal Manage 29:217–237
113. Pedersen G (1988) Three-dimensional wave patterns generated by moving disturbances at transcritical speeds. J Fluid Mech 196:39–63
114. Pedersen NF, Samuelsen MR, Welner D (1984) Soliton annihilation in the perturbed sine-Gordon system. Phys Rev B 30(7):4057–4059
115. Peregrine DH (1983) Wave jumps and caustics in the propagation of finite-amplitude water waves. J Fluid Mech 136:435–452
116. Perring JK, Skyrme THR (1962) A Model Unified Field Equation. Nucl Phys 31:550–555
117. Perroud PH (1957) The Solitary Wave Reflection Along a Straight Vertical Wall at Oblique Incidence. Univ. of Calif. Berkeley, IER Rept 99-3, pp 93
118. Peterson P, van Groesen E (2000) A direct and inverse problem for wave crests modelled by interactions of two solitons. Physica D 141:316–332
119. Peterson P, van Groesen E (2001) Sensitivity of the inverse wave crest problem. Wave Motion 34:391–399
120. Peterson P, Soomere T, Engelbrecht J, van Groesen E (2003) Soliton interaction as a possible model for extreme waves in shallow water. Nonlinear Process Geophys 10:503–510
121. Petviashvili VI (1980) Red Spot of Jupiter and the drift soliton in plasma. JETP Lett 32:619–622
122. Pigier C, Uzdin R, Carmon T, Segev M, Nepomnyaschchy A, Musslimani ZH (2001) Collisions between $(2+1)$D rotating propeller solitons. Opt Lett 26(20):1577–1579

123. Poladian L, Snyder AW, Mitchell DJ (1991) Spiraling spatial solitons. Opt Commun 85(1):59–62
124. Porubov AV, Tsuji H, Lavrenov IV, Oikawa M (2005) Formation of the rogue wave due to non-linear two-dimensional waves interaction. Wave Motion 42:202–210
125. Radhakrishnan R, Lakshmanan M, Hietarinta J (1997) Inelastic collision and switching of coupled bright solitons in optical fibers. Phys Rev E 56(2):2213–2216
126. Read PL, Hide R (1983) Long-lived eddies in the laboratory and in the atmosphere of Jupiter and Saturn. Nature 302(10):126–129
127. Rebbi C (1979) Solitons. Sci Am 240(2):76–91
128. Redekopp LG, Weidman PD (1978) Solitary Rossby waves in zonal shear flows and their interactions. J Atm Sci 35:790–804
129. Rotschild C, Alfassi B, Cohen O, Segev M (2006) Long-range interactions between optical solitons. Nat Phys 2(11):769–774
130. Rottman JW, Grimshaw R (2001) Atmospheric internal solitary waves. In: Grimshaw R (ed), Environmental Stratified Flows. Kluwer, Dordrecht, pp 91–129
131. Russell JS (1837) Applications and illustrations of the law of wave in the practical navigation of canals. Trans R Soc Edin 14:33–34
132. Russell JS (1844) Report on waves. In: Report of the 14th Meeting of the British Association for the Advancement of Science. Murray, York, pp 311–390
133. Salupere A, Peterson P, Engelbrecht J (2002) Long-time behaviour of soliton ensembles. Part I – Emergence of ensembles. Chaos Solitons Fractals 14(9):1413–1424
134. Salupere A, Peterson P, Engelbrecht J (2003) Long-time behaviour of soliton ensembles. Part II – Periodical patterns of trajectories. Chaos Solitons Fractals 15:29–40
135. Segev M (1998) Optical spatial solitons. Opt Quantum Electron 30(7–10):03–533
136. Segev M, Stegeman G (1998) Self-trapping of optical beams: Spatial solitons. Phys Today 51(8):42–48
137. Segev M, Crosignani B, Yariv A, Fischer B (1992) Spatial solitons in photorefractive media. Phys Rev Lett 68(7):923–926
138. Shalaby M, Barthelemy AJ (1992) Observation of the self-guided propagation of a dark and bright spatial soliton pair in a focusing nonlinear medium. IEEE J Quantum Electron 28(12):2736–2741
139. Shalaby M, Reynaud F, Barthelemy A (1992) Experimental-observation of spatial soliton-interactions with a pi-2 relative phase difference. Opt Lett 17(11):778–780
140. Shih MF, Segev M (1996) Incoherent collisions between two-dimensional bright steady-state photorefractive spatial screening solitons. Opt Lett 21(19):1538–1540
141. Shih MF, Segev M, Salamo G (1997) Three-dimensional spiraling of interacting spatial solitons. Phys Rev Lett 78(13):2551–2554
142. Snyder AW, Sheppard AP (1993) Collisions, steering, and guidance with spatial solitons. Opt Lett 18(7):482–484
143. Soljacic M, Sears S, Segev M (1998) Self-trapping of "necklace" beams in self-focusing Kerr media. Phys Rev Lett 81(22):4851–4854
144. Sommeria J, Nore C, Dumont T, Robert R (1991) Statistical theory of the Great Red Spot of Jupiter. C R Acad Sci II 312:999–1005
145. Soomere T (1992) Geometry of the double resonance of Rossby waves. Ann Geophys 10:741–748
146. Soomere T (2004) Interaction of Kadomtsev–Petviashvili solitons with unequal amplitudes. Phys Lett A 332:74–81
147. Soomere T (2005) Fast ferry traffic as a qualitatively new forcing factor of environmental processes in non-tidal sea areas: a case study in Tallinn Bay, Baltic Sea. Environ Fluid Mech 5:4 293–323
148. Soomere T (2006) Nonlinear ship wake waves as a model of rogue waves and a source of danger to the coastal environment: a review. Oceanologia 48(S):185–202
149. Soomere T (2007) Nonlinear components of ship wake waves. Appl Mech Rev 60(3):120–138
150. Soomere T, Engelbrecht J (2005) Extreme elevations and slopes of interacting solitons in shallow water. Wave Motion 41:179–192
151. Soomere T, Engelbrecht J (2006) Weakly two-dimensional interaction of solitons in shallow water. Eur J Mech B Fluids 25(5):636–648
152. Stegeman GI, Segev M (1999) Optical spatial solitons and their interactions: Universality and diversity. Science 286(5444):1518–1523. doi:10.1126/science.286.5444.1518
153. Stegner A, Zeitlin V (1996) Asymptotic expansions and monopolar solitary Rossby vortices in barotropic and two-layer model. Geophys Astrophys Fluid Dyn 83:159–195
154. Stegner A, Zeitlin V (1998) From quasi-geostrophic to strongly non-linear monopolar vortices in a paraboloidal shallow-water experiment. J Fluid Mech 356:1–24
155. Steiglitz K (2001) Time-gated Manakov spatial solitons are computationally. Phys Rev E 63(1):016608
156. Stepken A, Kaiser F, Belic MR, Krolikowski W (1998) Interaction of incoherent two-dimensional photorefractive solitons. Phys Rev E 58(4):R4112–R4115
157. Stern ME (1975) Minimal properties of planetary eddies. J Mar Res 33:1–13
158. Taijiri M, Maesono H (1997) Resonant interactions of drift vortex solitons in a convective motion of a plasma. Phys Rev E 55(3):3351–3357
159. Tanaka M (1993) Mach reflection of a large-amplitude solitary wave. J Fluid Mech 248:637–661
160. Tikhonenko V, Christou J, Luther-Davies B (1995) Spiraling bright spatial solitons formed by the breakup of an optical vortex in a saturable self-focusing medium. J Opt Soc Am B-Opt Phys 12(11):2046–2052
161. Tikhonenko V, Christou J, Luther-Davies B (1996) Three dimensional bright spatial soliton collision and fusion in a saturable nonlinear medium. Phys Rev Lett 76(15):2698–2701
162. Toffoli A, Lefevre JM, Bitner-Gregersen E, Monbaliu J (2005) Towards the identification of warning criteria: Analysis of a ship accident database. Appl Ocean Res 27(6):281–291
163. Torner L, Torres JP, Menyuk CR (1996) Fission and self-deflection of spatial solitons by cascading. Opt Lett 21(7):462–464
164. Torruellas WE, Wang Z, Hagan DJ, Vanstryland EW, Stegeman GI, Torner L, Menyuk CR (1995) Observation of 2-dimensional spatial solitary waves in a quadratic medium. Phys Rev Lett 74(25):5036–5039
165. Tratnik MV, Sipe JE (1988) Bound solitary waves in a birefringent optical fiber. Phys Rev A 38(4):2011–2017
166. Trillo S, Wabnitz S, Wright EM, Stegeman GI (1988) Optical solitary waves induced by cross-phase modulation. Opt Lett 13(10):871–873

167. Tsuji H, Oikawa M (2001) Oblique interaction of internal solitary waves in a two-layer fluid of infinite depth. Fluid Dyn Res 29:251–267
168. Tsuji H, Oikawa M (2004) Two-dimensional interaction of solitary waves in a modified Kadomtsev–Petviashvili equation. J Phys Soc Japan 73:3034–3043
169. Washimi H, Taniuti T (1966) Propagation of ion-acoustic solitary waves of small amplitude. Phys Rev Lett 17:996–998
170. Williams GP, Yamagata T (1984) Geostrophic regimes, intermediate solitary vortices and Jovian eddies. J Atm Sci 41:453–478
171. Yajima N, Oikawa M, Satsuma J (1978) Interaction of ion-acoustic solitons in three-dimensional space. J Phys Soc Japan 44:1711–1714
172. Yoon SB, Liu PLF (1989) Stem waves along breakwater. J Waterw Port Coast Ocean Eng – ASCE 115:635–648
173. Yue DK, Mei CC (1980) Forward diffrection of Stokes waves by a thin wedge. J Fluid Mech 99:33–52
174. Zabusky NJ, Kruskal MD (1965) Interaction on "solitons" in a collisionless plasma and the recurrence of initial states. Phys Rev Lett 15(6):240–243
175. Zakharov VE (1968) Stability of periodic waves of finite amplitude on the surface of a deep fluid. Zh Prikl Mekh Tekh Fiz 9:86–94; J Appl Mech Tech Phys 9:190–194
176. Zakharov VE, Shabat AB (1972) Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media. Sov Phys JETP 34:62–69
177. Zakharov VE, Manakov SV, Novikov SP, Pitaevsky LP (1980) Theory of solitons. Nauka, Moscow (in Russian) (English translation, 1984, Consultants Bureau, New York)
178. Ze F, Hershkowitz N, Chan C, Lonngren KE (1979) Inelastic collision of spherical ion-acoustic solitons. Phys Rev Lett 42(26):1747–1750

Solitons, Introduction to

Mohamed A. Helal
Department of Mathematics, Faculty of Science,
Cairo University, Giza, Egypt

The concept of solitons reflects one of the most important developments in science at the second half of the 20th century: the nonlinear description of the world. It is not easy to give a comprehensive and precise definition of a soliton. Frequently, a *soliton* is explained as a spatially localized wave in a medium that can interact strongly with other solitons but will afterwards regain its original form.

It is a nonlinear pulse-like wave that can exist in some nonlinear systems. The soliton wave can propagate without dispersing its energy over a large region of space; collision of two solitons leads to unchanged forms, solitons also exhibit particle-like properties.

The most remarkable property of solitons is that they do not disperse and thus conserve their form during propagation and collision. ▶ Solitons Interactions.

The nonlinear science has been growing for approximately fifty years. However, numerous nonlinear processes had been previously identified, but the nonlinear mathematical tools were not developed. The available tools were linear, and nonlinearities were avoided or treated as perturbations of linear theories.

The first experimental observation of a solitary wave was made in August 1834 by a Scottish engineer named John Scott Russell (1808–1882). Scott Russell reported his observation; to the Fourteenth Meeting of the British Association for Advancement of Science, held in 1844; in a long report entitled: "Report on waves" as follows:

> "*I was observing the motion of a boat which was rapidly drawn along a narrow channel by a pair of horses, when the boat suddenly stopped – not so the mass of water in the channel which it had put in motion; it accumulated round the prow of the vessel in a state of violent agitation, then suddenly leaving it behind, rolled forward with great velocity, assuming the form of a large solitary elevation, a rounded, smooth and well-defined heap of water, which continued its course along the channel apparently without change of form or diminution of speed. I followed it on horseback, and overtook it still rolling on at a rate of some eight or nine miles an hour, preserving its original figure some thirty feet long and a foot to a foot and a half in height. Its height gradually diminished, and after a chase of one or two miles I lost it in the windings of the channel. Such, in the month of August 1834, was my first chance interview with that singular and beautiful phenomenon which I have called the Wave of Translation*".

This hump-shape localized wave that propagates along one space-direction with undeformed shape has spectacular stability properties. John Scott Russell carried out many experiments to get the properties of this wave. ▶ Solitons: Historical and Physical Introduction.

A model equation describing the unidirectional propagation of long waves in water of relatively shallow depth with a localized solution (representing a single hump as discovered by Russell) was obtained by Korteweg and de Vries. This equation has become very famous and is now known as the Korteweg–de Vries equation or KdV equation. ▶ Water Waves and the Korteweg–de Vries Equation.

There exists a certain class of nonlinear partial differential equations that leads to solitons. The Korteweg–de Vries equation, Kadomtsev–Petviashvili (KP) equation, Klein–Gordon (KG) equation, Sine–Gordon (SG) equation, nonlinear Schrodinger (NLS) equation, Korteweg–de Vries Burger's (KdVB) equation, etc… are some known equations that lie in this specific class of this NLPDE. This class of nonlinear equations has a special type of the traveling wave solutions which are either solitary waves or solitons. ▶ Partial Differential Equations that Lead to Solitons.

The Korteweg–de Vries equation is a canonical equation in nonlinear wave physics demonstrating the existence of solitons and their elastic interactions. This KdV equation attracted many authors to publish an enormous number of publications. Different analytical methods for solving this well-known equation are presented in ▶ Korteweg–de Vries Equation (KdV), Different Analytical Methods for Solving the.

One of the most famous analytical methods for solving the KdV equation is the *Inverse Scattering Transformation*. This method was first introduced by the well-known scientists Gardener, Green, Kruskal, and Miura. In 1967, Gardner et al. published a magnificent and original paper on the analytical solution to the initial value problem due to a disturbance of finite amplitude in an infinite domain. This systematic method is not easy, and needs several long steps to get the complete solution. ▶ Inverse Scattering Transform and the Theory of Solitons.

There are various algebraic and geometric characteristics that the KdV equation possesses. More significantly, a lot of physically important solutions to the KdV equation can be presented explicitly through a simple, specific form, called *the Hirota bilinear form*. More complicated types of solutions leading to N-solitons are presented in ▶ Korteweg–de Vries Equation (KdV), History, Exact *N*-Soliton Solutions and Further Properties of the.

It was suitable to search for an easier method to solve this famous and well-known nonlinear partial differential equation (KdV). Many semi-analytical methods have been developed and used to solve the KdV equation. The most popular and famous semi-analytical method that was introduced to solve this nonlinear partial differential equation (KdV), was the *Adomian Decomposition Method*. ▶ Adomian Decomposition Method Applied to Non-linear Evolution Equations in Soliton Theory.

Other semi-analytical methods like the variational iteration method (VIM), homotopy analysis method (HAM), and homotopy perturbation method (HPM) are usable to solve these types of NLPDE such as the Korteweg–de Vries equation and even the modified Korteweg–de Vries equation (mKdV). ▶ Korteweg–de Vries Equation (KdV) and Modified Korteweg–de Vries Equations (mKdV), Semi–analytical Methods for Solving the.

It is indispensable to introduce other methods to illustrate the solution of this NLPDE. The numerical technique is a powerful tool which is always needed to present and compare the results in the easiest manner. There exist different numerical tools for solving the KdV and mKdV equations, and even for any evolution equations ▶ Korteweg–de Vries Equation (KdV), Some Numerical Methods for Solving the.

In hydrodynamical problems, the motion of incompressible inviscid fluid subject to a constant vertical gravitational force (where the fluid is bounded below by an impermeable bottom and above by a free surface) with some physical assumptions lead to the shallow water model. The shallow water wave equations are usually used in oceanography and atmospheric science. The shallow water approximation theory leads to the KdV equation. This means that the soliton solution as well as the solitary waves are the straight forward solutions for many physical problems. ▶ Shallow Water Waves and Solitary Waves.

The study of the water waves in the Pacific Ocean leads directly to a special type of long gravity waves named *tsunamis*. This word is a Japanese one, which means harbor's wave.

A tsunami is essentially a long wavelength water wave train or a series of waves generated in a body of water (mostly in oceans) that vertically displaces the water column. Earthquakes, landslides, volcanic eruptions, nuclear explosions and the impact of cosmic bodies can generate tsunamis. Tsunamis, as they approach coastlines, can rise enormously and savagely attack and inundate to cause huge damage to properties and cost thousands of lives.

The theoretical study of waves in oceanography is very complicated. The simplest mathematical model includes the KdV equation. This KdV represents the governing equation that lead to tsunamis. ▶ Solitons, Tsunamis and Oceanographical Applications of.

The rapid change in water density with depth is called *Pycnocline*. This happens in open or closed seas, as well as in oceans. In the frame of shallow water approximation theory that governs the oceans, and due to the pycnocline the nonlinear internal gravity waves are presented. They arise from perturbations to hydrostatic equilibrium, where balance is maintained between the force of gravity and the buoyant restoring force. An illustrative study of nonlinear waves (that was generated inside a stratified fluid occupying a semi infinite channel of finite and constant depth by a wave maker situated in motion at the finite extremity of the channel) is presented in this section. ▶ Non-linear Internal Waves.

The soliton perturbation theory is used to study solitons that are governed by the various nonlinear equations in the presence of the perturbation terms. ▶ Soliton Perturbation.

A *compacton* is a special solitary traveling wave that, unlike a soliton, does not have exponential tails.

Another expression and terminology called *compacton-like soliton* is a special wave solution which can be expressed by the squares of sinusoidal or cosinusoidal functions.

Soliton and compacton are two kinds of nonlinear waves. They play an indispensable and vital role in all branches of science and technology, and are used as constructive elements to formulate the complex dynamical behavior of wave systems throughout science: from hydrodynamics to nonlinear optics, from plasmas to shock waves, from tornados to the Great Red Spot of Jupiter, and from tsunamis to turbulence. More recently, soliton and compacton have been of key importance in the quantum fields and nanotechnology especially in nano-hydrodynamics. ► Solitons and Compactons.

Solitons, Tsunamis and Oceanographical Applications of

M. Lakshmanan
Center for Nonlinear Dynamics, Bharathidasan University, Tiruchirapalli, India

Article Outline

Glossary

Soliton A class of nonlinear dispersive wave equations in (1+1) dimensions having a delicate balance between dispersion and nonlinearity admit localized solitary waves which under interaction retain their shapes and speeds asymptotically. Such waves are called solitons because of their particle like elastic collision property. The systems include Korteweg–de Vries, nonlinear Schrödinger, sine-Gordon and other nonlinear evolution equations. Certain (2+1) dimensional generalizations of these systems also admit soliton solutions of different types (plane solitons, algebraically decaying lump solitons and exponentially decaying dromions).

Shallow and deep water waves Considering surface gravity waves in an ocean of depth h, they are called shallow-water waves if $h \ll \lambda$, where λ is the wavelength (or from a practical point of view if $h < 0.07\lambda$). In the linearized case, for shallow water waves the phase speed $c = \sqrt{gh}$, where g is the acceleration due to gravity. Water waves are classified as deep (practically) if $h > 0.28\lambda$ and the corresponding wave speed is given by $c = \sqrt{g/k}$, $k = \frac{2\pi}{\lambda}$.

Tsunami Tsunami is essentially a long wavelength water wave train, or a series of waves, generated in a body of water (mostly in oceans) that vertically displaces the water column. Earthquakes, landslides, volcanic eruptions, nuclear explosions and impact of cosmic bodies can generate tsunamis. Propagation of tsunamis is in many cases in the form of shallow water waves and sometimes can be of the form of solitary waves/solitons. Tsunamis as they approach coastlines can rise enormously and savagely attack and inundate to cause devastating damage to life and property.

Internal solitons Gravity waves can exist not only as surface waves but also as waves at the interface between two fluids of different density. While solitons were first recognized on the surface of water, the commonest ones in oceans actually happen underneath, as internal oceanic waves propagating on the pycnocline (the interface between density layers). Such waves occur in many seas around the globe, prominent among them being the Andaman and Sulu seas.

Rossby solitons Rossby waves are typical examples of quasigeostrophic dynamical response of rotating fluid systems, where long waves between layers of the atmosphere as in the case of the Great Red Spot of Jupiter or in the barotropic atmosphere are formed and may be associated with solitonic structures.

Bore solitons The classic bore (also called mascaret, poroca and aeger) arises generally in funnel shaped estuaries that amplify incoming tides, tsunamis or storm surges, the rapid rise propagating upstream against the flow of the river feeding the estuary. The profile depends on the Froude number, a dimensionless ratio of intertial and gravitational effects. Slower bores can take on oscillatory profile with a leading dispersive shockwave followed by a train of solitons.

Definition of the Subject

Surface and internal gravity waves arising in various oceanographic conditions are natural sources where one can identify/observe the generation, formation and propagation of solitary waves and solitons. Unlike the standard progressive waves of linear dispersive type, solitary waves are localized structures with long wavelengths and finite energies and propagate without change of speed or form and are patently nonlinear entities. The earliest scientifically recorded observation of a solitary wave was made by John Scott Russel in August 1834 in the Union Canal connecting the Scottish cities of Glasgow and Edinburgh. The theoretical formulation of the underlying phenomenon was provided by Korteweg and de Vries in 1895 who deduced the now famous Korteweg–de Vries (KdV) equation admitting solitary wave solutions. With the insightful numerical and analytical investigations of Martin Kruskal and coworkers in the 1960s the KdV solitary waves have been shown to possess the remarkable property that under collision they pass through each other without change of shape or speed except for a phase shift and so they are solitons. Since then a large class of soliton possessing nonlinear dispersive wave equations such as the sine-Gordon, modified KdV (MKdV) and nonlinear Schrödinger (NLS) equations occurring in a wide range of physical phenomena have been identified.

Several important oceanographic phenomena which correspond to nonlinear shallow water wave or deep water wave propagation have been identified/interpreted in terms of soliton propagation. These include tsunamis, especially earthquake induced ones like the 1960 Chilean or 2004 Indian Ocean earthquakes, internal solitary waves arising in stratified stable fluids such as the ones observed in Andaman or Sulu seas, Rossby waves including the Giant Red Spot of Jupiter and tidal bores occurring in estuaries of rivers. Detailed observations/laboratory experiments and theoretical formulations based on water wave equations resulting in the nonlinear evolution equations including KdV, Benjamin–Ono, Intermediate Long Wave (ILW), Kadomtsev–Petviashivili (KP), NLS, Davey–Stewartson (DS) and other equations clearly establish the relevance of soliton description in such oceanographic events.

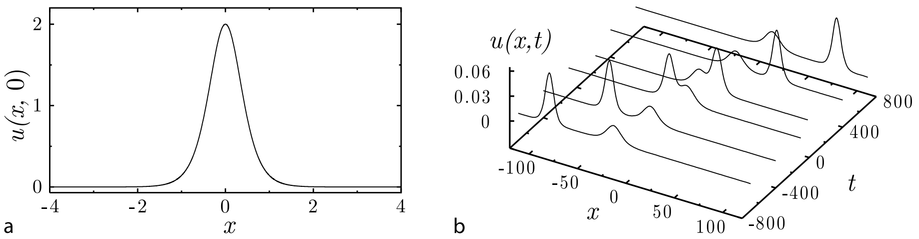
Introduction

Historically, the remarkable observation of John Scott Russel [1,2] of the solitary wave in the Union Canal connecting the cities of Edinburgh and Glasgow in the month of August 1834 may be considered as the precursor to the realization of solitons in many oceanographic phenomena. While riding on a horse back and observing the motion of boat drawn by a pair of horses which suddenly stopped but not so the mass of water it had set in motion, the wave (which he called the 'Great Wave of Translation') in the form of large solitary heap of water surged forward and travelled a long distance without *change of form or diminution of speed*. The wave observed by Scott Russel is nothing but a solitary wave having a remarkable staying power and a patently nonlinear entity. Korteweg and de Vries in 1895, starting from the basic equations of hydrodynamics and considering unidirectional shallow water wave propagation in rectangular channels, deduced [3] the now ubiquitous KdV equation as the underlying nonlinear evolution equation. It is a third order nonlinear partial differential equation in (1+1) dimensions with a delicate balance between dispersion and nonlinearity. It admits elliptic function cnoidal wave solutions and in a limiting form exact solitary wave solution of the type observed by John Scott Russel thereby vindicating his observations and putting to rest all the controversies surrounding them.

It was the many faceted numerical and analytical study of Martin Kruskal and coworkers [4,5] which firmly established by 1967 that the KdV solitary waves have the further remarkable feature that they are solitons having elastic collision property (Fig. 1).

It was proved decisively that the KdV solitary waves on collision pass through each other except for a finite phase shift, thereby retaining their forms and speeds asymptotically as in the case of particle like elastic collisions. The inverse scattering transform (IST) formalism developed for this purpose clearly shows that the KdV equation is a completely integrable infinite dimensional nonlinear Hamiltonian system and that it admits multisoliton solutions [6,7,8,9]. Since then a large class of nonlinear dispersive wave equations such as the sine-Gordon (s-G), modified Korteweg–de Vries (MKdV), NLS, etc. equations in (1+1) dimensions modeling varied physical phenomena have also been shown to be completely integrable soliton systems [6,7,8,9]. Interesting (2+1) dimensional versions of these systems such as Kadomtsev–Petviashvile (KP), Davey–Stewartson (DS) and Nizhnik–Novikov–Veselov (NNV) equations have also been shown to be integrable systems admitting basic nonlinear excitations such as line (plane) solitons, algebraically decaying lump solitons and exponentially localized dromion solutions [7,8].

It should be noted that not all solitary waves are solitons while the converse is always true. An example of a solitary wave which is not a soliton is the one which occurs in double-well $\lambda\phi^4$ wave equation which radiates energy on collision with another such wave. However even such solitary waves having finite energies are sometimes referred to as solitons in the condensed matter, particle physics and fluid dynamics literature because of their localized structure.



Solitons, Tsunamis and Oceanographical Applications of, Figure 1
KdV equation: a One soliton solution (solitary wave) at a fixed time, say $t = 0$, b Two soliton solution (depicting elastic collision)

As noted above solitary waves and solitons are abundant in oceanographic phenomena, especially involving shallow water wave and deep water wave propagations. However, these events are large scale phenomena very often difficult to measure experimentally, rely mostly on satellite and other indirect observations and so controversies and differing interpretations do exist in ascribing exact solitonic or solitary wave properties to these phenomena. Yet it is generally realized that many of these events are closely identifiable with solitonic structures. Some of these observations deserve special attention.

Tsunamis

When large scale earthquakes, especially of magnitude 8.0 in Richter scale and above, occur in seabeds at appropriate geological faults tsunamis are generated and can propagate over large distances as small amplitude and long wavelength structures when shallowness condition is satisfied. Though the tsunamis are hardly felt in the midsea, they take monstrous structures when they approach land masses due to conservation of energy. The powerful Chilean earthquake of 1960 [10] led to tsunamis which propagated for almost fifteen hours before striking Hawaii islands and a further seven hours later they struck the Japanese islands of Honshu and Hokkaido. More recently the devastating Sumatra–Andaman earthquake of 2004 in the Indian Ocean generated tsunamis which not only struck the coastlines of Asian countries including Indonesia, Thailand, India and Srilanka but propagated as far as Somalia and Kenya in Africa, killing more than a quarter million people. There is very good sense in ascribing soliton description to such tsunamis.

Internal Solitons

Peculiar striations, visible on satellite photographs of the surface of the Andaman and Sulu seas in the far east (and in many other oceans around the globe), have been interpreted as secondary phenomena accompanying the passage of “internal solitons”. These are solitary wavelike distortions of the boundary layer between the warm upper layer of sea water and cold lower depths. These internal solitons are travelling ridges of warm water, extending hundreds of meters down below the thermal boundary, and carry enormous energy. Osborne and Burch [11] investigated the underwater currents which were experienced by an oil rig in the Andaman sea, which was drilling at a depth of 3600 ft. One drilling rig was apparently spun through ninety degrees and moved one hundred feet by the passage of a soliton below.

Rossby Waves and Solitons

Rossby waves [12] are long waves between layers of the atmosphere, created by the rotation of the planet. In particular, in the atmosphere of a rotating planet, a fluid particle is endowed with a certain rotation rate, determined by its latitude. Consequently its motion in the north-south direction is constrained by the conservation of angular momentum as in the case of internal waves where gravity inhibits the vertical motion of a density stratified fluid. There is an analogy between internal waves and Rossby waves under suitable conditions. The KdV equation has been proposed as a model for the evolution of Rossby solitons [13] and NLS equation for the evolution of Rossby wave packets. The Great Red Spot of the planet of Jupiter is often associated with a Rossby soliton.

Bore Solitons

Rivers which flow to the open oceans are very often affected by tidal effects, tsunamis or storm surges. Tidal motions generate intensive water flows which can propagate upstream on tens of kilometers in the form of step-wise perturbation (hydraulic jumps) analogous to shock waves in acoustics. This phenomenon is known as a bore or mascaret (in French). Examples of such bores include the tidal bore of Seine river in France, Hooghly bore of the Ganges in India, the Amazon river bore in Brazil and Hangzhou bore in China. Bore disintegration into solitons is a possible phenomenon in such bores [14].

Besides these there are other possible oceanographic phenomena such as capillary wave solitons [15], resonant three-wave or four wave interaction solitons [16], etc., where also soliton picture is useful.

In this article we will first point out how in shallow channels the long wavelength wave propagation is described by KdV equation and its generalizations (Sect. "Shallow Water Waves and KdV Type Equations"). Then we will briefly point out how NLS family of equations arise naturally in the description of deep water waves (Sect. "Deep Water Waves and NLS Type Equations"). Based on these details, we will point how the soliton picture plays a very important role in the understanding of tsunami propagation (Sect. "Tsunamis as Solitons"), generation of internal solitons (Sect. "Internal Solitons"), formation of Rossby solitons (Sect. "Rossby Solitons") and disintegration of bores into solitons (Sect. "Bore Solitons").

Shallow Water Waves and KdV Type Equations

Kortweg and de Vries [3] considered the wave phenomenon underlying the observations of Scott Russel from first principles of fluid dynamics and deduced the KdV equation to describe the unidirectional shallow water wave propagation in one dimension.

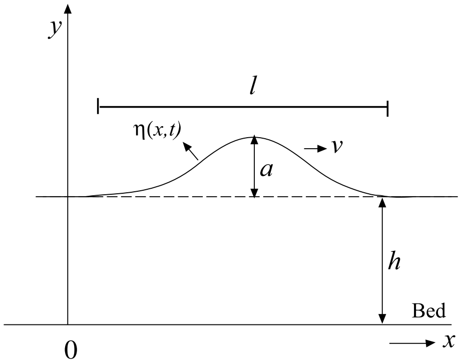
Consider the one-dimensional (x-direction) wave motion of an incompressible and inviscid fluid (water) in a shallow channel of height h, and of sufficient width with uniform cross-section leading to the formation of a solitary wave propagating under gravity. The effect of surface tension is assumed to be negligible. Let the length of the wave be l and the maximum value of its amplitude, $\eta(x,t)$, above the horizontal surface be a (see Fig. 2).

Then assuming $a \ll h$ (shallow water) and $h \ll l$ (long waves), one can introduce two natural small parameters into the problem $\epsilon = a/h$ and $\delta = h/l$. Then the analysis proceeds as follows [3,6].

Equation of Motion: KdV Equation

The fluid motion can be described by the velocity vector $\boldsymbol{V}(x,y,t) = u(x,y,t)\boldsymbol{i} + v(x,y,t)\boldsymbol{j}$, where $\boldsymbol{i}$ and $\boldsymbol{j}$ are the unit vectors along the horizontal and vertical directions, respectively. As the motion is irrotational, we have $\nabla \times \boldsymbol{V} = 0$. Consequently, we can introduce the velocity potential $\phi(x,y,t)$ by the relation $\boldsymbol{V} = \nabla\phi$.

Conservation of Density The system obviously admits the conservation law for the mass density $\rho(x,y,t)$ of the fluid, $\mathrm{d}\rho/\mathrm{d}t = \rho_t + \nabla\cdot(\rho\boldsymbol{V}) = 0$. As ρ is a constant, we have $\nabla\cdot\boldsymbol{V} = 0$. Consequently ϕ obeys the Laplace equation



Solitons, Tsunamis and Oceanographical Applications of, Figure 2
One-dimensional wave motion in a shallow channel

$$\nabla^2\phi(x,y,t) = 0\,. \tag{1}$$

Euler's Equation As the density of the fluid $\rho = \rho_0 =$ constant, using Newton's law for the rate of change of momentum, we can write $\mathrm{d}\boldsymbol{V}/\mathrm{d}t = \partial\boldsymbol{V}/\partial t + (\boldsymbol{V}\cdot\nabla)\,\boldsymbol{V} = -\frac{1}{\rho_0}\nabla p - g\boldsymbol{j}$, where $p = p(x,y,t)$ is the pressure at the point (x,y) and g is the acceleration due to gravity, which is acting vertically downwards (here $\boldsymbol{j}$ is the unit vector along the vertical direction). Since $\boldsymbol{V} = \nabla\phi$ we obtain (after one integration)

$$\phi_t + \frac{1}{2}(\nabla\phi)^2 + \frac{p}{\rho_0} + gy = 0\,. \tag{2}$$

Boundary Conditions The above two Eqs. (1) and (2) for the velocity potential $\phi(x,y,t)$ of the fluid have to be supplemented by appropriate boundary conditions, by taking into account the fact (see Fig. 2) that (a) the horizontal bed at $y = 0$ is hard and (b) the upper boundary $y = y(x,t)$ is a free surface. As a result
(a) the vertical velocity at $y = 0$ vanishes, $v(x,0,t) = 0$, which implies

$$\phi_y(x,0,t) = 0\,. \tag{3}$$

(b) As the upper boundary is free, let us specify it by $y = h + \eta(x,t)$ (see Fig. 2). Then at the point $x = x_1, y = y_1 \equiv y(x,t)$, we can write $\frac{\mathrm{d}y_1}{\mathrm{d}t} = \frac{\partial\eta}{\partial t} + \frac{\partial\eta}{\partial x}\cdot\frac{\mathrm{d}x_1}{\mathrm{d}t} = \eta_t + \eta_x u_1 = v_1$. Since $v_1 = \phi_{1y}$, $u_1 = \phi_{1x}$, we obtain

$$\phi_{1y} = \eta_t + \eta_x\phi_{1x}\,. \tag{4}$$

(c) Similarly at $y = y_1$, the pressure $p_1 = 0$. Then from (2), it follows that

$$u_{1t} + u_1 u_{1x} + v_1 v_{1x} + g\eta_x = 0\,. \tag{5}$$

Thus the motion of the surface of water wave is essentially specified by the Laplace Eq. (1) and Euler Eq. (2) along with one fixed boundary condition (3) and two *variable nonlinear* boundary conditions (4) and (5). One has to then solve the Laplace equation subject to these boundary conditions.

Taylor Expansion of $\boldsymbol{\phi(x,y,t)}$ in $\boldsymbol{y}$ Making use of the fact $\delta = h/l \ll 1$, $h \ll l$, we assume $y(= h + \eta(x,t))$ to be small to introduce the Taylor expansion

$$\phi(x,y,t) = \sum_{n=0}^{\infty} y^n \phi_n(x,t)\,. \tag{6}$$

Substituting the above series for ϕ into the Laplace Eq. (1), solving recursively for $\phi_n(x,t)$'s and making use of the boundary condition (4), $\phi_y(x,0,t) = 0$, one can show that

$$u_1 = \phi_{1x} = f - \frac{1}{2} y_1^2 f_{xx} + \text{ higher order in } \; y_1\,, \tag{7}$$

$$v_1 = \phi_{1y} = -y_1 f_x + \frac{1}{6} y_1^3 f_{xxx} + \text{ higher order in } \; y_1\,, \tag{8}$$

where $f = \partial\phi_0/\partial x$. We can then substitute these expressions into the nonlinear boundary conditions (4) and (5) to obtain equations for f and η.

Introduction of Small Parameters $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$ So far the analysis has not taken into account fully the shallow nature of the channel $(a/h = \epsilon \ll 1)$ and the solitary nature of the wave $(a/l = a/h \cdot h/l = \epsilon\delta \ll 1$, $\epsilon \ll 1$, $\delta \ll 1)$, which are essential to realize the Scott Russel phenomenon. For this purpose one can stretch the independent and dependent variables in the above equations through appropriate scale changes, but retaining the overall form of the equations. To realize this one can introduce the natural scale changes

$$x = lx'\,, \quad \eta = a\eta', \quad t = \frac{l}{c_0} t'\,, \tag{9}$$

where c_0 is a parameter to be determined. Then in order to retain the form of (7) and (8) we require

$$\begin{aligned} u_1 &= \epsilon c_0 u_1'\,, \quad v_1 = \epsilon\delta c_0 v_1'\,, \quad f = \epsilon c_0 f', \\ y_1 &= h + \eta(x,t) = h\left(1 + \epsilon\eta'\left(x', t'\right)\right)\,. \end{aligned} \tag{10}$$

Then

$$u_1' = f' - \frac{1}{2}\delta^2\left(1 + \epsilon\eta'\right)^2 f'_{x'x'} = f' - \frac{1}{2}\delta^2 f'_{x'x'}\,, \tag{11}$$

where we have omitted terms proportional to $\delta^2\epsilon$ as small compared to terms of the order δ^2. Similarly from (8), we obtain

$$v_1' = -\left(1 + \epsilon\eta'\right) f'_{x'} + \frac{1}{6}\delta^2 f'_{x'x'x'}\,. \tag{12}$$

Now considering the nonlinear boundary condition (4) in the form $v_1 = \eta_t + \eta_x u_1$, it can be rewritten as

$$\eta'_{t'} + f'_{x'} + \epsilon\eta' f'_{x'} + \epsilon f'\eta'_{x'} - \frac{1}{6}\delta^2 f'_{x'x'x'} = 0\,. \tag{13}$$

Similarly considering the other boundary condition (5) and making use of the above transformations, it can be rewritten, after neglecting terms of the order $\epsilon^2\delta^2$, as

$$f'_{t'} + \epsilon f' f'_{x'} + \frac{ga}{\epsilon c_0^2}\eta'_{x'} - \frac{1}{2}\delta^2 f'_{x'x't'} = 0\,. \tag{14}$$

Now choosing the arbitrary parameter c_0 as $c_0^2 = gh$ so that $\eta'_{x'}$ term is of order unity, (14) becomes

$$f'_{t'} + \eta'_{x'} + \epsilon f' f'_{x'} - \frac{1}{2}\delta^2 f'_{x'x't'} = 0\,. \tag{15}$$

(Note that $c_0 = \sqrt{gh}$ is nothing but the speed of the water wave in the linearized limit). Omitting the primes for convenience, the evolution equation for the amplitude of the wave and the function related to the velocity potential reads

$$\eta_t + f_x + \epsilon\eta f_x + \epsilon f\eta_x - \frac{1}{6}\delta^2 f_{xxx} = 0\,, \tag{16}$$

$$f_t + \eta_x + \epsilon f f_x - \frac{1}{2}\delta^2 f_{xxt} = 0\,. \tag{17}$$

Note that the small parameters ϵ and δ^2 have occurred in a natural way in (16), (17).

Perturbation Analysis Since the parameters ϵ and δ^2 are small in (16), (17), we can make a perturbation expansion of f in these parameters:

$$f = f^{(0)} + \epsilon f^{(1)} + \delta^2 f^{(2)} + \text{ higher order terms}\,, \tag{18}$$

where $f^{(i)}$, $i = 0, 1, 2, \ldots$ are functions of η and its spatial derivatives. Note that the above perturbation expansion is an asymptotic expansion. Substituting this into Eqs. (16) and (17) and regrouping and comparing different powers proportional to (ϵ, δ^2) and solving them successively one can obtain (see for example [6] for further details) in a self consistent way,

$$f^{(0)} = \eta, \quad f^{(1)} = -\frac{1}{4}\eta^2, \quad f^{(2)} = \frac{1}{3}\eta_{xx}. \tag{19}$$

Using these expressions into (18) and substituting it in (16) and (17), we ultimately obtain the KdV equation in the form

$$\eta_t + \eta_x + \frac{3}{2}\epsilon\eta\eta_x + \frac{\delta^2}{6}\eta_{xxx} = 0, \tag{20}$$

describing the unidirectional propagation of long wavelength shallow water waves.

The Standard (Contemporary) Form of KdV Equation

Finally, changing to a moving frame of reference, $\xi = x - t$, $\tau = t$, and introducing the new variables $u = (3\epsilon/2\delta^2)\eta$, $\tau' = (6/\delta^2)\tau$ and redefining the variables τ' as t and ξ as x, we finally arrive at the ubiquitous form of the KdV equation as

$$u_t + 6uu_x + u_{xxx} = 0. \tag{21}$$

The Korteweg–de Vries Eq. (21) admits cnoidal wave solution and in the limiting case solitary wave solution as well. This form can be easily obtained [6,17] by looking for a wave solution of the form $u = 2f(\xi)$, $\xi = (x - ct)$, and reducing the KdV equation into a third order nonlinear ordinary differential equation of the form

$$-c\frac{\partial f}{\partial \xi} + 12f\frac{\partial f}{\partial \xi} + \frac{\partial^3 f}{\partial \xi^3} = 0. \tag{22}$$

Integrating Eq. (22) twice and rearranging, we can obtain

$$\left(\frac{\partial f}{\partial \xi}\right)^2 = -4f^3 + cf^2 - 2df - 2b \equiv P(f), \tag{23}$$

where b and d are integration constants. Calling the three real roots of the cubic equation $P(f) = 0$ as α_1, α_2 and α_3 such that

$$\left(\frac{\partial f}{\partial \xi}\right)^2 = -4(f - \alpha_1)(f - \alpha_2)(f - \alpha_3), \tag{24}$$

the solution can be expressed in terms of the Jacobian elliptic function as

$$f(\xi) = f(x - ct) = \alpha_3 - (\alpha_3 - \alpha_2)\mathrm{sn}^2 \cdot \left[\sqrt{\alpha_3 - \alpha_1}\,(x - ct), m\right], \tag{25a}$$

where

$$(\alpha_1 + \alpha_2 + \alpha_3) = \frac{c}{4}, \quad m^2 = \frac{\alpha_3 - \alpha_2}{\alpha_3 - \alpha_1}, \quad \delta = \text{constant}. \tag{25b}$$

Here m is the modulus parameter of the Jacobian elliptic function. Eq. (25a) represents in fact the so called cnoidal wave (because of its elliptic function form). In the limiting case $m = 1$, the form (25a) reduces to

$$f = \alpha_2 + (\alpha_3 - \alpha_2)\ \mathrm{sech}^2\left[\sqrt{\alpha_3 - \alpha_1}\,(x - ct)\right]. \tag{26}$$

Choosing now $\alpha_1 = 0$, $\alpha_2 = 0$, and using (25b) we have

$$f = \frac{c}{4}\ \mathrm{sech}^2\left[\frac{\sqrt{c}}{2}\,(x - ct)\right]. \tag{27}$$

Then the solitary wave solution to the KdV Eq. (21) can be written in the form

$$u(x,t) = \frac{c}{2}\mathrm{sech}^2\frac{\sqrt{c}}{2}(x - ct + \delta), \quad \delta : \text{constant} \tag{28}$$

Note that the velocity of the solitary wave is directly proportional to the amplitude: larger the wave the higher is the speed. More importantly, the KdV solitary wave is a soliton: it retains its shape and speed upon collision with another solitary wave of different amplitude, except for a phase shift, see Fig. 1 [6,7,8]. In fact for an arbitrary initial condition, the solution of the Cauchy initial value problem consists of N-number of solitons of different amplitudes in the background of small amplitude dispersive waves. All these results ultimately lead to the result that the KdV equation is a completely integrable, infinite dimensional, nonlinear Hamiltonian system. It possesses [6,7,8]

(i) a Lax pair of linear differential operators and is solvable through the so called inverse scattering transform (IST) method,
(ii) infinite number of conservation laws and associated infinite number of involutive integrals of motion,
(iii) N-soliton solution,
(iv) Hirota bilinear form,
(v) Hamiltonian structure

and a host of other interesting properties (see for example [6,7,8,9]).

KdV Related Integrable and Nonintegrable NLEEs

Depending on the actual physical situation, the derivation of the shallow water wave equation can be suitably modified to obtain other forms of nonlinear dispersive wave equations in (1+1) dimensions as well as in (2+1) dimensions relevant for the present context. Without going into the actual derivations, some of the important equations possessing solitary waves are listed below [6,7,8,9].

1. Boussinesq equation [7]

$$u_t + uu_x + g\eta_x - \frac{1}{3}h^2 u_{txx} = 0, \tag{29a}$$

$$\eta_t + [u(h+\eta)]_x = 0 \tag{29b}$$

2. Benjamin–Bona–Mahoney (BBM) equation [18]

$$u_t + u_x + uu_x - u_{xxt} = 0 \tag{30}$$

3. Camassa–Holm equation [19]

$$u_t + 2\kappa u_x + 3uu_x - u_{xxt} = 3u_x u_{xx} + uu_{xxx} \tag{31}$$

4. Kadomtsev–Petviashville (KP) equation [7]

$$(u_t + 6uu_x + u_{xxx})_x + 3\sigma^2 u_{yy} = 0 \tag{32}$$

($\sigma^2 = -1$: KP-I, $\sigma^2 = +1$: KP-II).

In the derivation of the above equations, generally the bottom of water column or fluid bed is assumed to be flat. However in realistic situations the water depth varies as a function of the horizontal coordinates. In this situation, one often encounters inhomogeneous forms of the above wave equations. Typical example is the variable coefficient KdV equation [14]:

$$u_t + f(x,t)uu_x + g(x,t)u_{xxx} = 0, \tag{33}$$

where f and g are functions of x, t. More general forms can also be deduced depending upon the actual situations, see for example [14].

Deep Water Waves and NLS Type Equations

Deep water waves are strongly dispersive in contrast to the weakly dispersive nature of the shallow water waves (in the linear limit). Various authors (see for details [8]) have shown that nonlinear Schrödinger family of equations models the evolution of a packet of surface waves in this case. There are several oceanographic situations where such waves can arise [8]:

(i) A localized storm at sea can generate a wide spectrum of waves, which then propagates away from the source region in all horizontal directions. If the propagating waves have small amplitudes and encounter no wind away from the source region, these waves can eventually sort themselves into nearly one-dimensional packets of nearly monochromatic waves. For appropriately chosen scales, the underlying evolution of each of these packets can be shown to satisfy the nonlinear Schrödinger equation and its generalizations in (2+1) dimensions.

(ii) Nearly monochromatic, nearly one-dimensional waves can cover a broad range of surface waves in the sea that results due to a steady wind of long duration and fetch. Then the generalization of the NLS equation in (2+1) dimensions can describe the waves that result in when the wind stops.

In all the above situations one looks for the solution of the equations of motion (1)-(5) but generalized in three dimensions which consists mainly in the form of a small amplitude, nearly monochromatic, nearly one-dimensional wave train. Assuming that this wave train travels in the x-direction with a mean wave number $\boldsymbol{\kappa} = (k, l)$ with a characteristic amplitude 'a' of the disturbance and a characteristic variation δk in k, one can deduce the NLS equation in (2+1) dimensions under the following conditions:

(i) small amplitudes such that $\hat{\epsilon} = \epsilon\delta \equiv \kappa a \ll 1$
(ii) slowly varying modulations, $\frac{\delta k}{\kappa} \ll 1$
(iii) nearly one dimensional waves, $\frac{|l|}{k} \ll 1$
(iv) balance of all three effects, $\frac{\delta k}{\kappa} = \frac{|l|}{k} \approx O(\hat{\epsilon})$
(v) for finite and deep water waves, $(kh)^2 \gg \hat{\epsilon}$

In the lowest order approximation (linear approximation) of water waves, the prediction is that a localized initial state will generally evolve into wave packets with a dominant wave number $\boldsymbol{\kappa}$ and corresponding frequence ω, given by the dispersion relation $\omega = (g\kappa + \sigma\kappa^3)\tanh\kappa h$, $\kappa = |\boldsymbol{\kappa}| = \sqrt{k^2 + l^2}$ within which each wave propagates with the phase speed $c = \omega/k$, while the envelope propagates with the group velocity $c_g = \mathrm{d}\omega/\mathrm{d}\kappa$. After a sufficiently long time the wave packet tends to disperse around the dominant wave number.

This tendency for dispersion can be offset by cumulative nonlinear effects. In the absence of surface tension, the outcome for unidirectional waves can be shown to be describable by the NLS equation. If the surface wave in the lowest order is $\phi \approx A\exp i(kx - \omega t)$+c.c, where ϕ is the velocity potential, then to leading order the wave amplitude evolves as

$$i(A_t + c_g A_x) + \frac{1}{2}\lambda A_{xx} + \mu|A|^2 A = 0. \tag{34}$$

The coefficients here are given by

$$\lambda = \frac{\partial^2 \omega}{\partial k^2}\,,$$
$$\mu = -\frac{\omega k^2}{16 S^4}(8C^2 S^2 + g - 2T^2) + \frac{\omega}{8C^2 S^2}\frac{(2\omega C^2 + k c_g)^2}{(gh - c_g^2)}\,, \tag{35}$$

where $C = \cosh(kh)$, $S = \sinh(kh)$, $T = \frac{S}{C}$. Equation (34) has been obtained originally by Zakharov in 1968 for deep water waves [20] and by Hasimoto and Ono for waves of finite depth in 1972 [21].

The NLS equation is also a soliton possessing integrable system and is solvable by the IST method [6,7,8,9]. For $\lambda > 0$ (focusing case), the envelope solitary (soliton) wave solution (also called bright solitons in the optical physics context) is given by

$$A(x,t) = a\, \text{sech}\gamma(x - c_g t)\exp(-i\Omega t)\,, \tag{36}$$

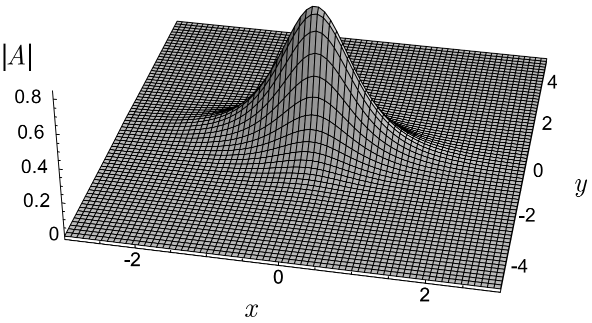
where $\mu a^2 = \lambda \gamma^2$, $\Omega = +\frac{1}{2}\mu a^2$.

When the effects of modulation in the transverse y-direction are taken into account, so that the wave amplitude is now given by $A(x,y,t)$, the NLS equation is replaced by the Benney–Roskes system [22] also popularly known as the Davey–Stewartson equations [23],

$$i(A_t + c_g A_x) + \frac{1}{2}\lambda A_{xx} + \frac{1}{2}\delta A_{yy} + \mu|A|^2 A + UA = 0\,, \tag{37a}$$

$$\alpha U_{xx} + U_{yy} + \beta(|A|^2)_{yy} = 0\,, \tag{37b}$$

where $\delta = \frac{c_g}{k}$, $\alpha = 1 - \left(\frac{c_g^2}{gh}\right)$, $gh\beta = \frac{\omega}{8C^2S^2}(2\omega C^2 + kc_g)^2$. Here $U(x,y,t)$ is the wave induced mean flow. In the deep water wave limit, $kh \to \infty$, and $U \to 0$, $\beta \to 0$ and one has the nonintegrable (2+1) dimensional NLS equation. On the other hand in the shallow water limit, one has the integrable Davey–Stewartson (DS) equations. For details see [7,8]. The DS-I equation admits algebraically decaying lump solitons and exponentially decaying dromions [24] besides the standard line solitons for appropriate choices of parameters. A typical dromion solution of DS-I equation is shown in Fig. 3.

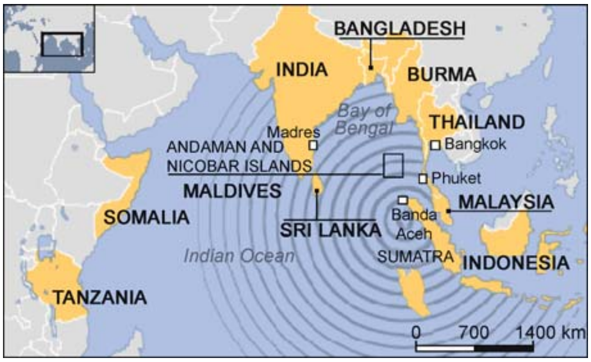


Solitons, Tsunamis and Oceanographical Applications of, Figure 3
Exponentially localized dromion solution of the Davey–Stewartson equation at a fixed time ($t = 0$) for suitable choice of parameters

Tsunamis as Solitons

The term ‘tsunami’ (tsu:harbour, nami:wave in Japanese) which was perhaps an unknown word even for scientists in countries such as India, Srilanka, Thailand, etc. till recently has become a house-hold word since that fateful morning of December 26, 2004. When a powerful earthquake of magnitude 9.1–9.3 on the Richter scale, epicentered off the coast of Sumatra, Indonesia, struck at 07:58:53, local time, described as the 2004 Indian Ocean earthquake or Sumatra–Andaman earthquake (Fig. 4), it triggered a series of devastating tsunamis as high as 30 meters that spread throughout the Indian Ocean killing about 275,000 people and inundating coastal communities across South and Southeast Asia, including parts of Indonesia, Srilanka, India and Thailand and even reaching as far as the east coast of Africa [25]. The catastrophe is considered to be one of the deadliest disasters in modern history.

Since this earthquake and consequent tsunamis, several other earthquakes of smaller and larger magnitudes keep occurring off the coast of Indonesia. Even as late as July 17, 2006 an earthquake of magnitude 7.7 on the Richter scale struck off the town of Pandering at 15.19 lo-



Solitons, Tsunamis and Oceanographical Applications of, Figure 4
26 December 2004 Indian Ocean tsunami (adapted from the website www.blogaid.org.uk with the courtesy of Andy Budd)

cal time and set off a tsunami of 2m high which had killed more than 300 people.

These tsunamis, which can become monstrous tidal waves when they approach coastline, are essentially triggered due to the sudden vertical rise of the seabed by several meters (when earthquake occurs) which displaces massive volume of water. The tsunamis behave very differently in deep water than in shallow water as pointed out below. By no means the tsunami of 2004 and later ones are exceptional; More than two hundred tsunamis have been recorded in scientific literature since ancient times. The most notable earlier one is the tsunami triggered by the powerful earthquake (9.6 magnitude) off southern Chile on May 22, 1960 [10] which traveled almost 22 hours before striking Japanese islands.

It is clear from the above events that the tsunami waves are fairly permanent and powerful ones, having the capacity to travel extraordinary distances without practically diminishing in size or speed. In this sense they seem to have considerable resemblance to shallow water nonlinear dispersive waves of KdV type, particularly solitary waves and solitons. It is then conceivable that tsunami dynamics has close connection with soliton dynamics.

Basics of Tsunami Waves

As noted above tsunami waves of the type described above are essentially triggered by massive earthquakes which lead to vertical displacement of a large volume of water. Other possible reasons also exist for the formation and propagation of tsunami waves: underwater nuclear explosion, larger meteorites falling into the sea, volcano explosions, rockslides, etc. But the most predominant cause of tsunamis appear to be large earthquakes as in the case of the Sumatra–Andaman earthquake of 2004. Then there are three major aspects associated with the tsunami dynamics [26]:

1. Generation of tsunamis
2. Propagation of tsunamis
3. Tsunami run up and inundation

There exist rather successful models to approach the generation aspects of tsunamis when they occur due to the earthquakes [27]. Using the available seismic data it is possible to reconstruct the permanent deformation of the sea bottom due to earthquakes and simple models have been developed (see for example, [28]). Similarly the tsunami run up and inundation problems [29] are extremely complex and they require detailed critical study from a practical point of view in order to save structures and lives when a tsunami strikes.

However, here we will be more concerned with the propagation of tsunami waves and their possible relation to wave propagation associated with nonlinear dispersive waves in shallow waters. In order to appreciate such a possible connection, we first look at the typical characteristic properties of tsunami waves as in the case of 2004 Indian Ocean tsunami waves or 1960 Chilean tsunamis.

The Indian Ocean Tsunami of 2004

Considering the Indian Ocean 2004 tsunami, satellite observations after a couple of hours after the earthquake establish an amplitude of approximately 60 cms in the open ocean for the waves. The estimated typical wavelength is about 200 kms [30]. The maximum water depth h is between 1 and 4 kms. Consequently, one can identify in an average sense the following small parameters (ϵ and δ^2) of roughly equal magnitude:

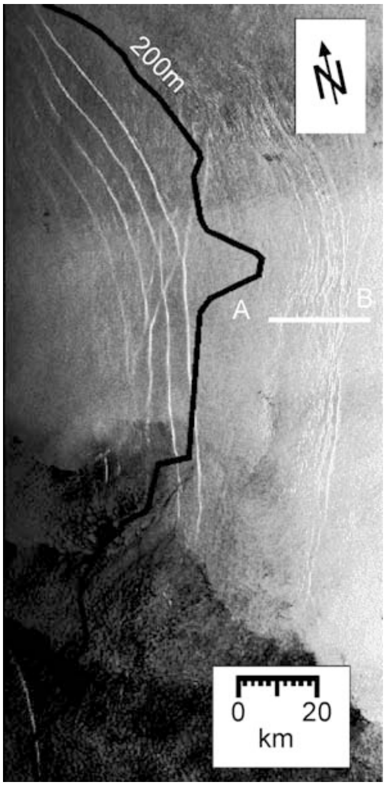
$$\epsilon = \frac{a}{h} \approx 10^{-4} \ll 1, \quad \delta^2 = \frac{h^2}{l^2} \approx 10^{-4} \ll 1 \tag{38}$$

As a consequence, it is possible that a nonlinear shallow water wave theory where dispersion (KdV equation) also plays an important role (as discussed in Sect. "Shallow Water Waves and KdV Type Equations") has considerable relevance [26]. However, we also wish to point out here that there are other points of view: Constantin and Johnson [31] estimate $\epsilon \approx 0.002$ and $\delta \approx 0.04$ and conclude that for both nonlinearity and dispersion to become significant the quantity $\delta\epsilon^{-3/2}\times$wavelength estimated as 90,000 kms is too large and shallow water equations with variable depth (without dispersion) should be used. However, it appears that these estimates can vary over a rather wide range and with suitable estimates it is possible that the range of 10,000–20,000 kms could be also possible ranges and hence taking into account the fact that both the Indian Ocean 2004 and Chilean 1960 tsunamis have traveled over 10 hours or more (in certain directions) before encountering land mass appears to allow for the possibility of nonlinear dispersive waves as relevant features for the phenomena. Segur [32] has argued that in the 2004 tsunamis, the propagation distances from the epicenter of the earthquake to India, Srilanka, or Thailand were too short for KdV dynamics to develop. In the same way one can argue that the waves that hit Somalia and Kenya in the east coast of Africa (or as in the case of Chilean earthquake see also [32]) have traveled sufficiently long distance for KdV dynamics to become important [33]. In any case one can conclude that at least for long distance tsunami propagation solitary wave and soliton picture of KdV like equations become very relevant.

Internal Solitons

For a long time seafarers passing through the Strait of Malacca on their journeys between India and the Far East have noticed that in the Andaman sea, between Nicobar islands and the north east coast of Sumatra, often bands of strongly increased surface roughness (ripplings or bands of choppy water) occur [11,34]. Similar observations have been reported in other seas around the globe from time to time. In recent times there has been considerable progress in understanding these kind of observations in terms of internal solitons in the oceans [7]. These studies have been greatly facilitated by photographs taken from satellites and space-crafts orbiting the earth, for example by synthetic aperture radar (SAR) images of ERS-1/2 satellites [34,35].

Peculiar striations of 100 km long, separated by 6 to 15 km and grouped in packets of 4 to 8, visible on satellite photographs (see Fig. 5) of the surface of the Andaman and Sulu seas in the Far East, have been interpreted as secondary phenomena accompanying the passage of 'internal solitons', which are solitary wavelike distortions of the boundary layer between warm upper layer of sea water and cold lower depths. These internal solitons are traveling edges of warm water, extending hundreds of meters down below the thermal boundary [7]. They carry enormous energy with them which is perhaps the reason for unusually strong underwater currents experienced by deep-sea drilling rigs. Thus these internal solitons are potentially hazardous to sub-sea oil and gas explorations. The ability to predict them can improve substantially the cost effectiveness and safety of offshore drilling.



Solitons, Tsunamis and Oceanographical Applications of, Figure 5

SAR image of a 200 km × 200 km large section of the Andaman Sea acquired by the ERS-2 satellite on April 15, 1996 showing sea surface manifestations of two internal solitary wave packets, see [34]. Figure reproduced from ESA website www.earth.esa.int/workshops/ers97/papers/alpers3 with the courtesy of European Space Agency and W. Alpers

A systematic study of the underwater currents experienced by an oil rig in the Andaman sea which was drilling at a depth of 3600 ft was carried out by Osborne and Burch in 1980 [11]. They spent four days measuring underwater currents and temperatures. The striations seen on satellite photographs turned out to be kilometer-wide bands of extremely choppy water, stretching from horizon to horizon, followed by about two kilometers of water "as smooth as a millpond". These bands of agitated water are called "tide rips", they arose in packets of 4 to 8, spaced about 5 to 10 km apart (they reached the research vessel at approximately hourly intervals) and this pattern was repeated with the regularity of tidal phenomenon.

As described in [7], Osborne and Burch found that the amplitude of each succeeding soliton was less than the previous one, is precisely what is expected for solitons (note that the velocity of a solitary wave solution of KdV equation increases with amplitude, vide Eq. (22)). Thus if a number of solitons are generated together, then we expect them eventually to be arranged in an ordered sequence of decreasing amplitude. From the spacing between successive waves in a packet and the rate of separation calculated from the KdV equation, Osborne and Burch were able to estimate the distance the packet had traveled from its source and thus identify possible source regions [7]. They concluded that the solitons are generated by tidal currents off northern Sumatra or between the islands of the Nicobar chain that extends beyond it and that their observations have good general agreement with the predictions for internal solitons as given by the KdV equation. Numerous recent observations and predictions of solitons in the Andaman sea have clearly established that it is a site where extraordinarily large internal solitons are encountered [34,35].

Further, Apel and Holbrook [36] undertook a detailed study of internal waves in the Sulu sea. Satellite photographs had suggested that the source of these waves was near the southern end of the Sulu sea and their research ship followed one wave packet for more than 250 miles over a period of two days – an extraordinary coherent phenomenon [7]. These internal solitons travel at speeds of about 8 kilometers per hour (5 miles per hour), with amplitude of about 100 meters and wavelength of about 1700 meters.

Similar observations elsewhere have confirmed the presence of internal solitons in oceans including the strait of Messina, the strait of Gibraltar, off the western side of Baja California, the Gulf of California, the Archipelago of La Maddalena and the Georgia strait [7]. There has also been a number of experimental studies of internal solitons in laboratory tanks in the last few decades [37]. These experiments provide detailed quantitative information usually unavailable in the field conditions, and are also an efficient tool for verifying various theoretical models.

As a theoretical formulation of internal solitons [7], consider two incompressible, immiscible fluids, with densities ρ_1 and ρ_2 and depths h_1 and h_2 respectively such that the total depth $h = h_1 + h_2$. Let the lighter fluid of height h_1 be lying over a heavier fluid of height h_2, in a constant gravitational field (Fig. 6). The lower fluid is assumed to rest on a horizontal impermeable bed, and the upper fluid is bounded by a free surface.

Then as in Sect. "Shallow Water Waves and KdV Type Equations", we denote the characteristic amplitude of wave by 'a' and the characteristic wavelength $l = k^{-1}$. Then the various nonlinear wave equations to describe the formation of internal solitons follow by suitable modification of the formulation in Sect. "Shallow Water Waves and KdV Type Equations", assuming viscous effects to be negligible. Each of these equations is completely integrable and admits soliton solutions [7].

(a) KdV equation (Eq. (21)) follows when

(i) the waves are of long wavelength $\delta = \frac{h}{l} \ll 1$,
(ii) the amplitude of the waves are small, $\epsilon = \frac{a}{h} \ll 1$, and
(iii) the two effects are comparable $\delta^2 = O(\epsilon)$

(b) Intermediate-Long-Wave (ILW) equation [38]

$$u_t + u_x + 2uu_x + Tu_{xx} = 0, \tag{39}$$

where Tu is the singular integral operator

$$(Tf)(x) = \frac{1}{2l} \fint_{-\infty}^{\infty} \coth\left\{\frac{\pi}{2l}(y-x)\right\} f(y)\mathrm{d}y \tag{40}$$

with $\fint_{-\infty}^{\infty}$ the Cauchy principal value integral is obtained under the assumption that

(a) there is a thin (upper) layer, $\epsilon = \frac{h_1}{h_2} \ll 1$,
(b) the amplitude of the waves is small, $a \ll h_1$,
(c) the above two effects balance, $\frac{a}{h_1} = O(\epsilon)$,
(d) the characteristic wavelength is comparable to the total depth of the fluid, $l = kh = O(1)$ and
(e) the waves are long waves in comparison with the thin layer, $kh_1 \ll 1$.

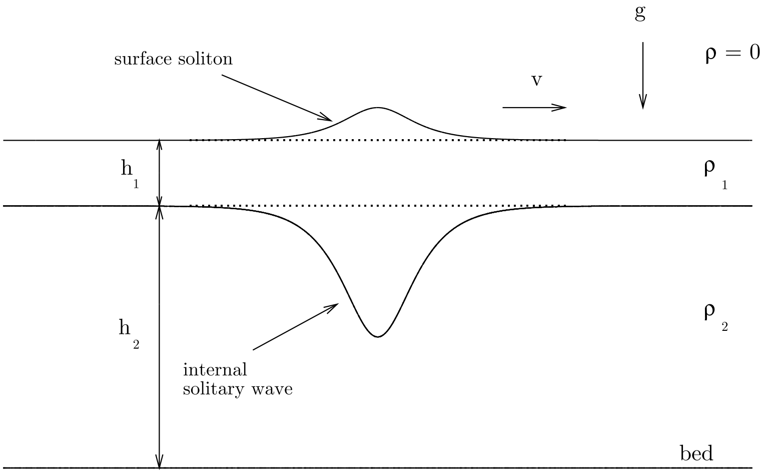
(c) Benjamin–Ono equation [39]

$$u_t + 2uu_x + Hu_{xx} = 0, \tag{41}$$

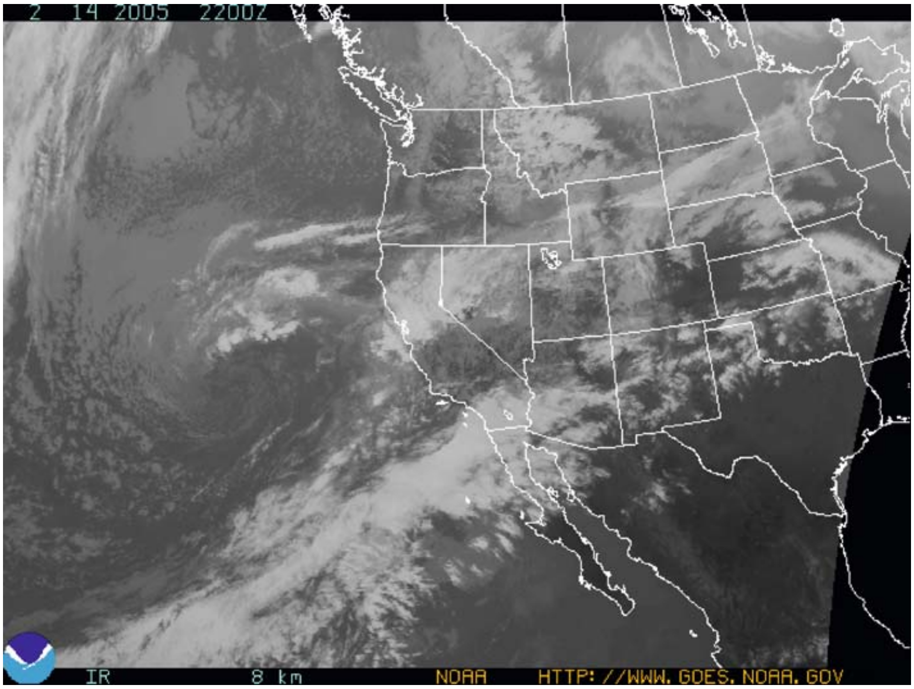
where Hu is the Hilbert transform

$$(Hf)(x) = \frac{1}{\pi} \fint_{-\infty}^{\infty} \frac{f(y)}{y-x}\mathrm{d}y, \tag{42}$$

is obtained under the assumption



Solitons, Tsunamis and Oceanographical Applications of, Figure 6
Formation of internal soliton (note that under suitable conditions small amplitude surface soliton can also be formed)



Solitons, Tsunamis and Oceanographical Applications of, Figure 7

A coupled high/low pressure systems in the form of a Rossby soliton formed off the coast of California/Washington on Valentine's Day 2005. The low pressure front hovers over Los Angeles dumping 30 inches of rain on the city in two weeks. The high pressure system lingers off the coast of Washington state providing unseasonally warm and sunny weather. This particular Rossby soliton proved exceptionally stable because of its remarkable symmetry. In an accompanying animated gif in this website one can watch the very interesting phenomenon as the jet stream splits around the soliton suctioning warm wet air from the off the coast of Mexico to Arizona, leaving behind a welcome drenching of rain. The figure and caption have been adapted from the website http://mathpost.la.asu.edu/~rubio/rossby_soliton/rs.html with the courtesy of the National Oceanic and Atmospheric Administration (NOAA) and Antonio Rubio

(a) there is a thin (upper) layer $h_1 \ll h_2$,
(b) the waves are long waves in comparison with the thin layer, $kh_1 \ll 1$,
(c) the waves are short in comparison with the total depth of the fluid, $kh \gg 1$ and
(d) the amplitude of the waves is small, $a \ll h_1$.

It may be noted that in the shallow water limit, as $\delta \to 0$, the ILW equation reduces to the KdV equation, while the Benjamin–Ono equation reduces to it in the deep water wave limit as $\delta \to \infty$. Each of these equations have their own ranges of validity and admit solitary wave and soliton solutions to represent internal solitons of the oceans.

Rossby Solitons

The atmospheric dynamics is an important subject of human concern as we live within the atmosphere and are continuously affected by the weather and its rather complex behavior. The motion of the atmosphere is intimately connected with that of the ocean with which it exchanges fluxes of momentum, heat and moisture [40]. Then the atmospheric dynamics is dominated by the rotation of the earth and the vertical density stratification of the surrounding medium, leading to newer effects. Similar effects are also present in other planetary dynamics as well.

In the atmosphere of a rotating planet, a fluid particle is endowed with a certain rotation rate (Coriolis frequency), determined by its latitude. Consequently its motion in the north-south direction is inhibited by conservation of angular momentum. The large scale atmospheric waves caused by the variation of the Coriolis frequency with latitude are called *Rossby waves*. In Sect. "Internal Solitons" we saw that KdV equation and its modifications model internal waves. Since there is a resemblance between internal waves and Rossby waves, it is expected that KdV like equations can model Rossby waves as well [8]. Under the simplest assumptions like long waves, incompressible fluid, β-plane approximations, etc. Benney [41] had derived the KdV equation as a model for Rossby waves in the presence of east-west zonal flow. Boyd [42] had shown that long, weakly nonlinear, equatorial Rossby waves are governed either by KdV or MKdV equation. Recently it has been shown by using the eight years of Topex/Poseidon altimeter observations [43] that a detailed characterization of major components of Pacific dynamics confirms the presence of equatorial Rossby solitons.

Also observational and numerical studies of propagation of nonlinear Rossby wave packets in the barotropic atmosphere by Lee and Held [44,45] have established their presence, notably in the Northern Hemisphere as storm tracks, but more clearly in the Southern Hemisphere (see Fig. 7). They also found that the wavepackets both in the real atmosphere and in the numerical models behave like the envelope solitons of the nonlinear Schrödinger equation (Fig. 7).

An interesting application of KdV equation to describe Rossby waves is the conjecture of Maxworthy and Redekopp [46] that the planet Jupiter's Great Red Spot might be a solitary Rossby wave. Photographs taken by the spacecraft Voyager of the cloud pattern show that the atmospheric motion on Jupiter is dominated by a number of east-west zonal currents, corresponding to the jet streams of the earth's atmosphere [8]. Several oval-shaped spots are also seen, including the prominent Great Red Spot in the southern hemisphere. The latter one has been seen approximately this latitude for hundreds of years and maintains its form despite interactions with other atmospheric objects. One possible explanation is that the Red Spot is a solitary wave/soliton of the KdV equation, deduced from the quasigeostrophic form of potential vorticity equation for an incompressible fluid. A second test of the model is the combined effect of interaction of the Red Spot and Hollow of the Jupiter atmosphere on the South Tropical Disturbance which occurred early in the 20th century and lasted several decades. Maxworthy and Redekopp [46] have interpreted this interaction as that of two soliton collision of KdV equation, with the required phase shift [6,7,8].

The above connection between the KdV solitary wave and the rotating planet may be seen more concretely by the following argument as detailed in [8]. One can start with the quasigeostrophic form of the potential vorticity equation for an incompressible fluid in the form

$$\left\{\left(\frac{\partial}{\partial t}+v\frac{\partial}{\partial x}\right)+\epsilon\left(\frac{\partial\psi}{\partial y}\right)\left(\frac{\partial}{\partial x}-\frac{\partial\psi}{\partial x}\frac{\partial}{\partial y}\right)\right\}$$
$$\left\{\mu^2\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial}{\partial z}\left(K^2\frac{\partial}{\partial z}\right)\right\}\psi+(\beta-U'')\frac{\partial\psi}{\partial x}=0, \tag{43}$$

where x, y and z represent the east, north and vertical directions, while the function U is related to the horizontal stream function Φ as $\Phi(x,y,z,t) = \int_{y'}^{y} U(\eta)d\eta + \epsilon\psi(x,y,z,t)$ in terms of a zonal shear flow and a perturbation. In (43), in the β-plane approximation, the Coriolis parameter is given as $f = 2\Omega\sin\theta_0 + \beta y$ and the function $K(z)$ is given by $K(z) = 2\Omega\sin\theta_0 l_2/N(z)d$, where $N(z)$ is the Brunt–Väisälä frequency and l_2 is the characteristic length scales in the north-south direction while d is the length scale in the vertical direction. Note that K compares the effects of rotation and density of variation. Then $\mu = l_2/l_1$ represents the ratio of the length scales in the north-south and east-west directions.

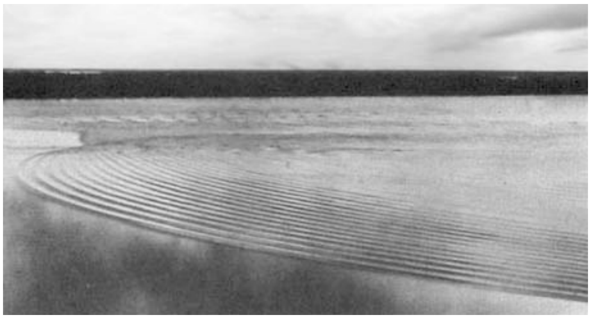
In the linear ($\epsilon \to 0$), long wave ($\mu \to 0$) limit, the potential function ψ has the form $\psi = \sum_n A_n(x - c_n t)\phi_n(y)p_n(z)$, where c_n is deduced by solving two related eigenvalue problems [8], $(K^2 p'_n)' + k_n^2 p_n = 0$, $p_n(0) = p_n(1) = 0$ and $\phi''_n - k_n^2\phi_n + [(\beta - U'')/(U - c_n)]\phi_n = 0$, $\phi_n(y_s) = \phi_n(y_N) = 0$. If the various modes are separable on a short time scale, one can then deduce an evolution equation for the individual modes, by eliminating secular terms at higher order in the expansion. Depending on the nature and existence of a stable density of stratification, which is characterized by the function $N(z)$ mentioned above, either KdV or mKdV equation can be deduced for a given mode and thereby establishing the soliton connection in the present problem.

In spite of the complexity of the phenomenon underlying Rossby solitons there is clear evidence of the significance of solitonic picture.

Bore Solitons

Rivers which flow into the open oceans are usually affected by tidal flows, tsunami or storm surge. For a typical estuary as one moves towards the mouth of the river, the depth increases and width decreases. When a tidal wave, tsunami or storm surge hits such an estuary, it can be seen as a hydraulic jump (step-wise perturbations like a shock-wave) in the water height and speed which will propagate upstream [47]. Far less dangerous but very similar is the bore (mascaret in French), a tidal wave which can propagate in a river for considerable distances.

Typical examples of bores occur in Seine river in France and the Hooghli river in West Bengal in India. A bore existed in the Seine river upto 1960 and disappeared when the river was dredged. The tide amplitude here is one of the largest in the world. The Hughli river is a branch of the Ganges that flows through Kolkata where a bore of 1m is present, essentially due to the shallowness of the river. Another interesting example is that at the time



Solitons, Tsunamis and Oceanographical Applications of, Figure 8
Tidal bore at the mouth of the Araguari River in Brazil. The bore is undular with over 20 waves visible. (Adapted from the book "Gravity Currents in the Environment and the Laboratory" by John E. Simpson, Cambridge University Press with the courtesy of D.K. Lynch, John E. Simpson and Cambridge University Press)

of the 1983 Japan sea tsunami, waves in the form of a bore ascended many rivers [48]. In some cases, bores had the form of one initial wave with a train of smaller waves and in other cases only a step with flat water surface behind was observed (Fig. 8). The Hangzhou bore in China is a tourist attraction. Other well known bores occur in the Amazon in Brazil and in Australia.

Another interesting situation where bore solitons were observed was in the International H_2O Project (IHOP), as a density current (such as cold air from thunderstorm) intrudes into a fluid of lesser density that occurs beneath a low level inversion in the atmosphere. A spectacular bore and its evolution into a beautiful amplitude-ordered train of solitary waves were observed and sampled during the early morning of 20 June 2002 by the Leandre-II abroad the P-3 aircraft. The origin of this bore was traceable to a propagating cold outflow boundary from a mesoscale convective system in extreme western Kansas [49].

In the process of propagation the bore undergoes dissipation, dispersive disintegration, enhancement due to decrease of the river width and depth, influence of nonlinear effects and so on. The profile depends on the Froude number which is a dimensionless ratio of inertial and gravitational effects. Theoretical models have been developed to study these effects based on KdV and its generalizations [14]. For example, bore disintegration into solitons in channel of constant parameters can be studied in signal coordinates in terms of the KdV like equation for the perturbation of water surface,

$$\eta_x + \frac{1}{c_0}(1-\alpha\eta)\eta_t - \beta\eta_{ttt} = 0\,, \qquad (44)$$

where $c_0 = \sqrt{gh}$, $\alpha = 3/2h$, $\beta = h^2/6c_0^3$, h being the depth of the river, with the bore represented by a Heaviside step function as the boundary condition at $x = 0$. Other effects then can be incorporated into a variable KdV equation of the form (26).

Future Directions

We have indicated a few of the most important oceanographical applications of solitons including tsunamis, internal solitons, Rossby solitons and bore solitons. There are other interesting phenomena like capillary wave solitons [15], resonant three and four wave interaction solitons [16], etc. which are also of considerable interest depending on whether the wave propagation corresponds to shallow, intermediate or deep waters. Whatever be the situation, it is clear that experimental observations as well as their theoretical formulation and understandings are highly challenging complex nonlinear evolutionary problems. The main reason is that the phenomena are essentially large scale events and detailed experimental observations require considerable planning, funding, technology and manpower. Often satellite remote sensing measurements need to be carried out as in the case of internal solitons and Rossby solitons. Events like tsunami propagation are rare and time available for making careful measurements are limited and heavily dependent on satellite imaging, warning systems and after event measurements. Consequently developing and testing theoretical models are extremely hazardous and difficult. Yet the basic modeling in terms of solitary wave/soliton possessing nonlinear dispersive wave equations such as the KdV and NLS family of equations and their generalizations present fascinating possibilities to understand these large scale complex phenomena and opens up possibilities of prediction. Further understanding of such nonlinear evolution equations, both integrable and nonintegrable systems particularly in higher dimensions, can help to understand the various phenomena clearly and provide means of predicting events like tsunamis and damages which occur due to internal solitons and bores. Detailed experimental observations can also help in this regard. It is clear that what has been understood so far is only the qualitative aspects of these phenomena and much more intensive work is needed to understand the quantitative aspects to predict them.

Bibliography

Primary Literature

1. Russel JS (1844) Reports on Waves. 14th meeting of the British Association for Advancement of Science. John Murray, London, pp 311–390

2. Bullough RK (1988) The Wave Par Excellence. The solitary progressive great wave of equilibrium of fluid: An early history of the solitary wave. In: Lakshmanan M (ed) Solitons: Introduction and Applications. Springer, Berlin
3. Korteweg DJ, de Vries G (1895) On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves. Philos Mag 39:422–443
4. Zabusky NJ, Kruskal MD (1965) Interactions of solitons in a collisionless plasma and recurrence of initial states. Phys Rev Lett 15:240–243
5. Gardner CS, Greene JM, Kruskal MD, Miura RM (1967) Method for solving the Korteweg–de Vries equation. Phys Rev Lett 19:1095–97
6. Lakshmanan M, Rajasekar S (2003) Nonlinear Dynamics: Integrability and Chaos. Springer, Berlin
7. Ablowitz MJ, Clarkson PA (1991) Solitons, Nonlinear Evolution Equations and Inverse Scattering. Cambridge University Press, Cambridge
8. Ablowitz MJ, Segur H (1981) Solitons and Inverse Scattering Transform. Society for Industrial and Applied Mathematics, Philadelphia
9. Scott AC (1999) Nonlinear Science: Emergence and Dynamics of Coherent Structures. Oxford University Press, New York
10. Dudley WC, Miu L (1988) Tsunami! University of Hawaii Press, Honololu
11. Osborne AR, Burch TL (1980) Internal solitons in the Andaman Sea. Science 258:451–460
12. Rossby GG (1939) Relation between variations in the intensity of the zonal circulation of the atmosphere. J Mar Res 2:38–55
13. Redekopp L (1977) On the theory of solitary Rossby waves. J Fluid Mech 82:725–745
14. Caputo JG, Stepanyants YA (2003) Bore formation and disintegration into solitons in shallow inhomogeneous channels. Nonlinear Process Geophys 10:407–424
15. Longuet–Higgins MS (1993) Capillary gravity waves of solitary type and envelope solitons in deep water. J Fluid Mech 252:703–711
16. Philips OM (1974) Nonlinear dispersive waves. Ann Rev Fluid Mech 6:93–110
17. Helal MA, Molines JM (1981) Nonlinear internal waves in shallow water. A theoretical and experimental study. Tellus 33:488–504
18. Benjamin TB, Bona JL, Mahoney JJ (1972) Model equations for long waves in nonlinear dispersive systems. Philos Trans A Royal Soc 272:47–78
19. Camassa R, Holm D (1992) An integrable shallow water equation with peaked solitons. Phys Rev Lett 71:1661–64
20. Zakharov VE (1968) Stability of periodic waves of finite amplitude on the surface of a deep fluid. J Appl Mech Tech Phys 2:190–194
21. Hasimoto H, Ono H (1972) Nonlinear modulation of gravity waves. J Phys Soc Jpn 33:805–811
22. Benney DJ, Roskes G (1969) Wave instabilities. Stud Appl Math 48:377–385
23. Davey A, Stewartson K (1974) On three dimensional packets of surface waves, Proc Royal Soc Lond A 338:101–110
24. Fokas AS, Santini PM (1990) Dromions and a boundary value problem for the Davey–Stewartson I equation. Physica D 44:99–130
25. Kundu A (ed) (2007) Tsunami and Nonlinear Waves. Springer, Berlin
26. Dias F, Dutykh D (2007) Dynamics of tsunami waves. In: Ibrahimbegovic A, Kozar I (eds) Extreme man-made and natural hazards in dynamics of structures. NATO security through Science Series. Springer, Berlin, pp 201–224
27. Dutykh D, Dias F (2007) Water waves generated by a moving bottom. In: Kundu A (ed) Tsunami and Nonlinear Waves. Springer, Berlin, pp 65–94
28. Okada Y (1992) Internal deformation due to shear and tensile faults in a half space. Bull Seism Soc Am 82:1018–1040
29. Carrier GF, Wu TT, Yeh H (2003) Tsunami runup and drawdown on a plane beach. J Fluid Mech 475:79–99
30. Banerjee P, Politz FF, Burgman R (2005) The size and duration of the Sumatra–Andaman earthquake from far-field static offsets. Science 308:1769–1772
31. Constantin A, Johnson RS (2006) Modelling tsunamis. J Phys A39:L215-L217
32. Segur H (2007) Waves in shallow water, with emphasis on the tsunami of (2004). In: Kundu A (ed) Tsunami and Nonlinear Waves. Springer, Berlin, pp 3–29
33. Lakshmanan M (2007) Integrable nonlinear wave equations and possible connections to tsunami dynamics. In: Kundu A (ed) Tsunami and Nonlinear Waves. Springer, Berlin, pp 31–49
34. Alpers W, Wang-Chen H, Cook L (1997) Observation of internal waves in the Andaman Sea by ERS SAR. IEEE Int 4:1518–1520
35. Hyder P, Jeans DRG, Cauqull E, Nerzic R (2005) Observations and predictability of internal solitons in the northern Andaman Sea. Appl Ocean Res 27:1–11
36. Apel JR, Holbroook JR (1980) The Sulu sea internal soliton experiment, 1. Background and overview. EOS Trans AGU 61:1009
37. Ostrovsky LA, Stepanyants YA (2005) Internal solitons in laboratory experiments: Comparison with theoretical models. Chaos 15(1–28):037111
38. Joseph RI (1977) Solitary waves in a finite depth fluid. J Phys A10:L225-L227
39. Benjamin TB (1967) Internal waves of permanent form in fluids of great depth. J Fluid Mech 29:559–592
40. Kundu PK, Cohen IM (2002) Fluid Mechanics, Second Edition. Academic Press, San Diego
41. Benney DJ (1966) Long nonlinear waves in fluid flows. Stud Appl Math 45:52–63
42. Boyd JP (1980) Equatorial solitary waves. Part I: Rossby solitons. J Phys Oceanogr 10:1699–1717
43. Susanto RD, Zheng Q, Xiao-Hai Y (1998) Complex singular value decomposition analysis of equatorial waves in the Pacific observed by TOPEX/Poseidon altimeter. J Atmospheric Ocean Technol 15:764–774
44. Lee S, Held I (1993) Baroclinic wave packets in models and observations. J Atmospheric Sci 50:1413–1428
45. Tan B (1996) Collision interactions of envelope Rossby solitons in a baratropic atmosphere. J Atmospheric Sci 53:1604–1616
46. Maxworthy T, Redekopp LG (1976) Theory of the Great Red Spot and other observed features of the Jovian atmosphere. Icarus 29:261–271
47. Caputo JG, Stepanyants YA (2007) Tsunami surge in a river: a hydraulic jump in an inhomogeneous channel. In: Kundu A (ed) Tsunami and Nonlinear Waves. Springer, Berlin, pp 97–112
48. Tsuji T, Yanuma T, Murata I, Fujiwara C (1991) Tsunami ascending in rivers as an undular bore. Nat Hazard 4:257–266
49. Koch SE, Pagowski M, Wilson JW, Fabry F, Flamant C, Feltz W, Schwemmer G, Geerts B (2005) The structure and dynamics of

atmospheric bores and solitons as determined from remote sensing and modelling experiments during IHOP, AMS 32nd Conference on Radar Meteorology, Report JP6J.4

Books and Reviews

Lamb H (1932) Hydrodynamics. Dover, New York
Miles JW (1980) Solitary waves. Ann Rev Fluid Mech 12:11–43
Stoker JJ (1957) Water Waves. Interscience, New York
Dauxois T, Peyrard M (2006) Physics of Solitons. Cambridge University Press, Cambridge
Johnson RS (1997) An Introduction to the Mathematical Theory of Water Waves. Cambridge University Press, Cambridge
Hammack JL (1973) A note on tsunamis: their generation and propagation in an ocean of uniform depth. J Fluid Mech 60:769–800
Drazin PG, Johnson RS (1989) Solitons: An Introduction. Cambridge University Press, Cambridge
Mei CC (1983) The Applied Dynamics of Ocean Surface Waves. Wiley, New York
Scott AC et al (1973) The soliton: a new concept in applied science. Proc IEEE 61:1443–83
Scott AC (ed) (2005) Encyclopedia of Nonlinear Science. Routledge, New York
Sulem C, Sulem P (1999) The Nonlinear Schrödinger Equation. Springer, Berlin
Helfrich KR, Melville WK (2006) Long nonlinear internal waves. Ann Rev Fluid Mech 38:395–425
Bourgault D, Richards C (2007) A laboratory experiment on internal solitary waves. Am J Phys 75:666–670

Space Plasmas, Dynamical Complexity in

TOM CHANG
Kavli Institute for Astrophysics and Space Research,
Massachusetts Institute of Technology, Cambridge, USA

Article Outline

Glossary

Alfvén wave Transverse magnetohydrodynamic wave that propagates along the magnetic field line direction in an electrically conducting fluid.

Anomalous resistivity Nonclassical resistivity in plasma dynamics.

Astronomical unit (AU) Unit of length nearly equal to the semi-major axis of Earth's orbit around the Sun.

Auroral electrojet index (AE) Empirical index that measures magnetic perturbations in the auroral zone. It provides a gauge for the horizontal current strength in the auroral ionosphere.

Auroral zone Region where auroral activities are observed.

Castaing distribution A convolution of Gaussians with variances distributed according to a log-normal distribution.

Chaos Nonsteady, nonperiodic, complex dynamical motion.

Coarse-grained dissipation Apparent dissipation due to interactions of coherent structures.

Coarse-graining Replacement of a detailed description with a lower-resolution coarse-grained model.

Coherent structure Large scale structure of nonlinearly interacting dynamical systems.

Correlation function Ensemble average of the correlation of fluctuations.

Criticality Behavior of extended system at which scale-invariance prevails.

Crossover Transition from one critical state to another or multitudes of critical states; or, transition from one fractal state to another or multitudes of fractal states.

Current sheet Sheet-like region with strong current density usually generated by strong local magnetic shear.

Cusp Funnel shaped region dividing the sunward and tailward sides of the magnetic field lines of the Earth.

Drift Alfvén vortex Field-aligned coherent Alfvénic vortex structure that drifts across the magnetic field.

Dynamical complexity Complex stochastic behavior of numerous nonlinearly interacting coherent structures of various sizes.

Dynamical renormalization group Group generated by coarse-graining scale transformations of a nonlinear dynamical system.

Extended similarity scaling (ESS) Scale invariant property among the structure or partition functions of a fluctuating event.

Exponent relation Algebraic relation among scaling exponents.

Extreme value theory Statistical theory dealing with extreme deviations from the mean.

Flatness (kurtosis) A normalized scale dependent measure related to the fourth order moment of the scale dependent coarse-grained probability distribution. It may be evaluated by taking the ensemble average of the square of the normalized power using the wavelet transform.

Forced and/or self-organized criticality (FSOC) Organization of a dynamical system into a scale invariant critical state with or without some tuning or forcing.

Fractal Fluctuations that exhibit generalized dimension of an irrational number.

Gaussian distribution Normal distribution.

Generalized dimension Dimension of the measure of fractal characteristics of a fluctuating event based on the limiting characteristic of the logarithm of the partition function at small scales.

Helicity Extent to which the magnetic field wraps around itself.

Heliosphere Magnetic bubble containing our solar system, including the solar wind.

Hysteresis History dependence of physical system that dissipates energy.

Intermittency Random fluctuations that exhibit non-self-similar characteristics.

Irrelevant parameter Parameter that becomes irrelevant under repeated coarse-graining transformations.

Ionosphere Upper most portion of atmosphere that is ionized and connected to the magnetosphere.

Local intermittency measure, Lim Normalized power in wavelet transform that is intensity, scale and location dependent.

Low-dimensional chaos Chaotic motion characterized by a small number of parameters.

Lu/Klimas model One-dimensional dynamical model with anomalous diffusion involving hysteresis, which exhibits scale invariant stochastic processes.

Magnetic reconfiguration (and/or reconnection) Reconfiguration of magnetic topology, which may or may not involve the phenomenon of magnetic reconnection of field lines.

Magnetohydrodynamics Mathematical description of electrically conducting fluids moving within electromagnetic environments.

Magnetosphere Region of space around the Earth (or any astrophysical object), that is controlled by its magnetic field.

Magnetotail Tail region (away from the Sun) of the magnetosphere of the Earth or any object in the heliosphere.

Multifractals Fluctuations characterized by multitudes of fractal dimensions that are not linearly related.

P-model Two-scale cascade model with "p" as the parameter that characterizes the fragmentation probability of the cascading process of intermittency.

Partition function of order q qth moment-order of the segmental coarse-grained probabilities.

PDF Probability distribution function.

Plasma resonance Site at which a particular plasma propagation mode vanishes.

Rank-ordered multifractal analysis, ROMA Analysis of multifractal characteristics based on the ordering of fractal dimensions in terms of scaled sizes of the intermittent fluctuations.

Relevant parameter Parameter that becomes more and more relevant under repeated coarse-graining transformations.

Response function Ensemble average of the response of the system due to fluctuations.

Scale invariants Quantities that do not vary under coarse-graining scale transformations.

Singularity spectrum $f(\alpha)$ Continuum distribution of differential measures with singularity index α.

Self-organized criticality (SOC) Auto-organization of a dynamical system into a scale-invariant critical state without significant tuning.

Solar wind Stream of charged particles in interplanetary space ejected from the upper atmosphere of the Sun.

Space plasma Ionized medium in the space environment.

Structure function of order q qth order moment of the probability distribution function of scale-dependent, coarse-grained fluctuations. It may be evaluated in terms of the ensemble average of the qth power of the scale-dependent fluctuations.

Taylor hypothesis Hypothesis that interprets temporal fluctuations as spatial fluctuations when the transit time of the fluctuations is much less than the characteristic evolution time. Also called Taylor's "frozen-in" hypothesis.

Turbulence Chaotic stochastic flow of a continuum or discrete medium with infinite number of degrees of freedom.

Wavelet transform Transform generated by a basis set of wave-packet-like functions, which provides information of intensity, location and scale of a fluctuating quantity.

Whistler wave Right-hand polarized electromagnetic wave in a plasma with frequency well below the electron cyclotron frequency and above the ion cyclotron frequency. The high frequency whistler waves travel faster than the low frequency ones along the magnetic field lines in the plasma medium.

Definition of the Subject

Commonly called the fourth state of matter, plasmas make up 99 per cent of the Universe. They are the major constituents of the solar-terrestrial, interstellar, and intergalactic media – generally in the form of ionized gases. According to Tonks [118] who co-authored the celebrated Tonks and Langmuir [119] paper on oscillations in ionized gases, Langmuir [70] coined the terminology "plasma" to describe, in today's language, a quasineutral mixture – probably because of the likeness between the electrons and ions in an ionized medium and the red and white corpuscles in a blood plasma [99]. Recent observations, particularly in-situ measurements in the heliosphere and solar-terrestrial regions, have indicated that space plasmas commonly exhibit random, intermittent, and anisotropic fluctuating characteristics at all spatiotemporal scales while interacting with the electromagnetic environment. Such observations have led theoreticians and experimentalists alike to regard the stochastic behaviors of space plasmas as prime candidates of naturally occurring examples of dynamical complexity.

By definition, "dynamical complexity" is a phenomenon exhibited by a nonlinearly interacting dynamical system within which multitudes of different sizes of large scale "coherent structures" are formed, resulting in a complicated global nonlinear stochastic behavior for the dynamical system – vastly different from that could be surmised from the original dynamical equations.

The main purpose of this article is to demonstrate the prevalence of dynamical complexity in space plasmas and to indicate the various statistical methods of analyzing this kind of dynamical processes as well as to discuss theoretical and numerical methods that can provide the basic understanding of such processes.

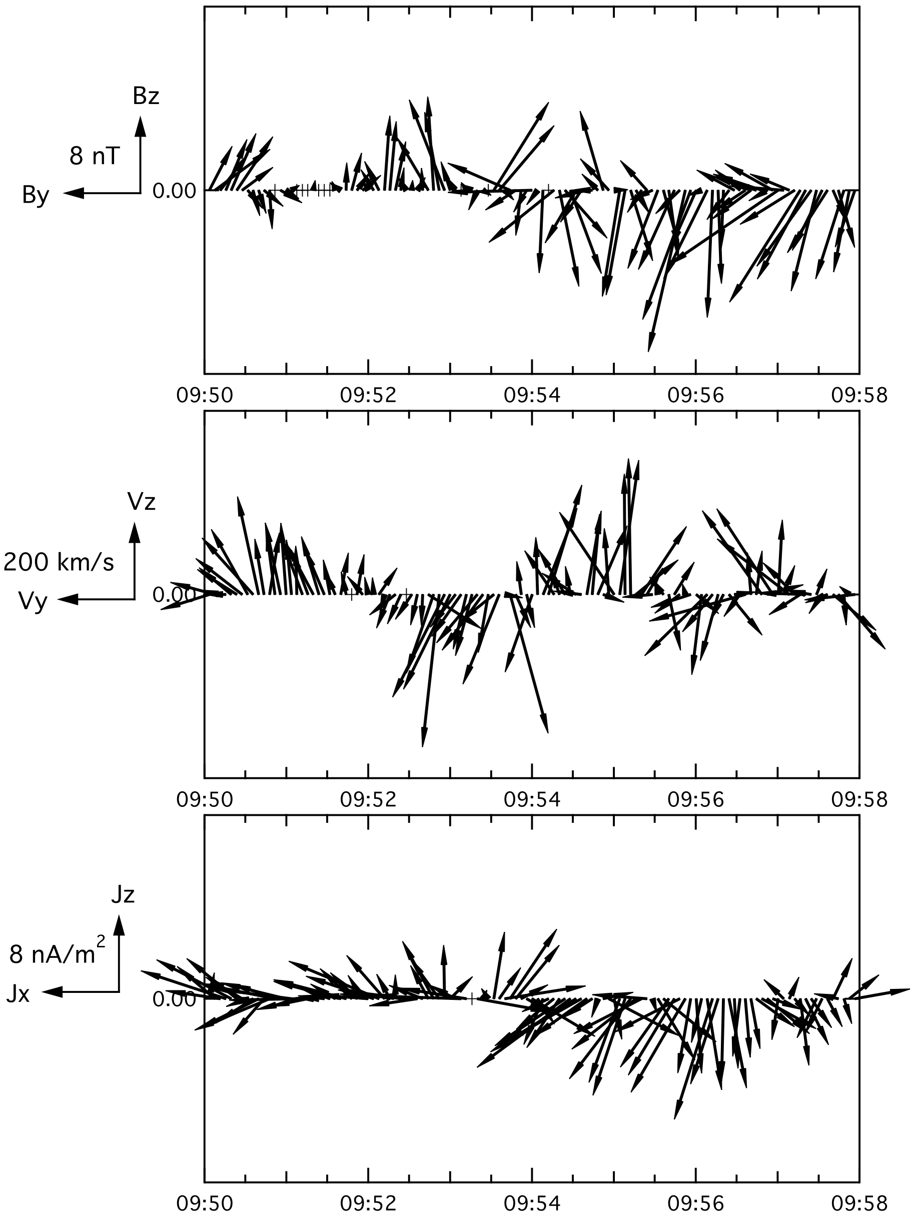
Introduction

Traditionally, analyzes of space plasma processes are based on fluid or kinetic formulations. Among the fluid descriptions, the simplest is that based on the so-called one-component magnetohydrodynamic (MHD) equations, which includes the equations of motion, continuity, and energy conservation, the Ohm's law and the Maxwell equations for a single conducting fluid medium (see, e. g. [7]). Associated with such a formulation are the concepts of magnetic field lines, streamlines and other continuum variables such as velocity, electric and magnetic fields, as well as plasma, charge and current densities. The dynamical state of the system is understood in terms of a topology generally characterized by the relatively smooth or piecewise continuous variations of such entities with space and time.

Most of the observed space plasma processes, however, generally exhibit discernible turbulent fluctuations of such quantities. The standard approach to the theoretical analyzes of dynamically fluctuating states is based on the concepts of linear instabilities, nonlinear growths and interactions of the wave modes. Although the basic governing equations generally contain strong nonlinearities, one is led to believe that turbulent motions may be understood by expressing the fluctuations in Fourier modes (plane waves) and then considering the interactions among these nonlocal modes by requiring them to satisfy the "basic" equations. This procedure produced such intractable complications that have led to decades of futile search for a workable theory of "turbulence".

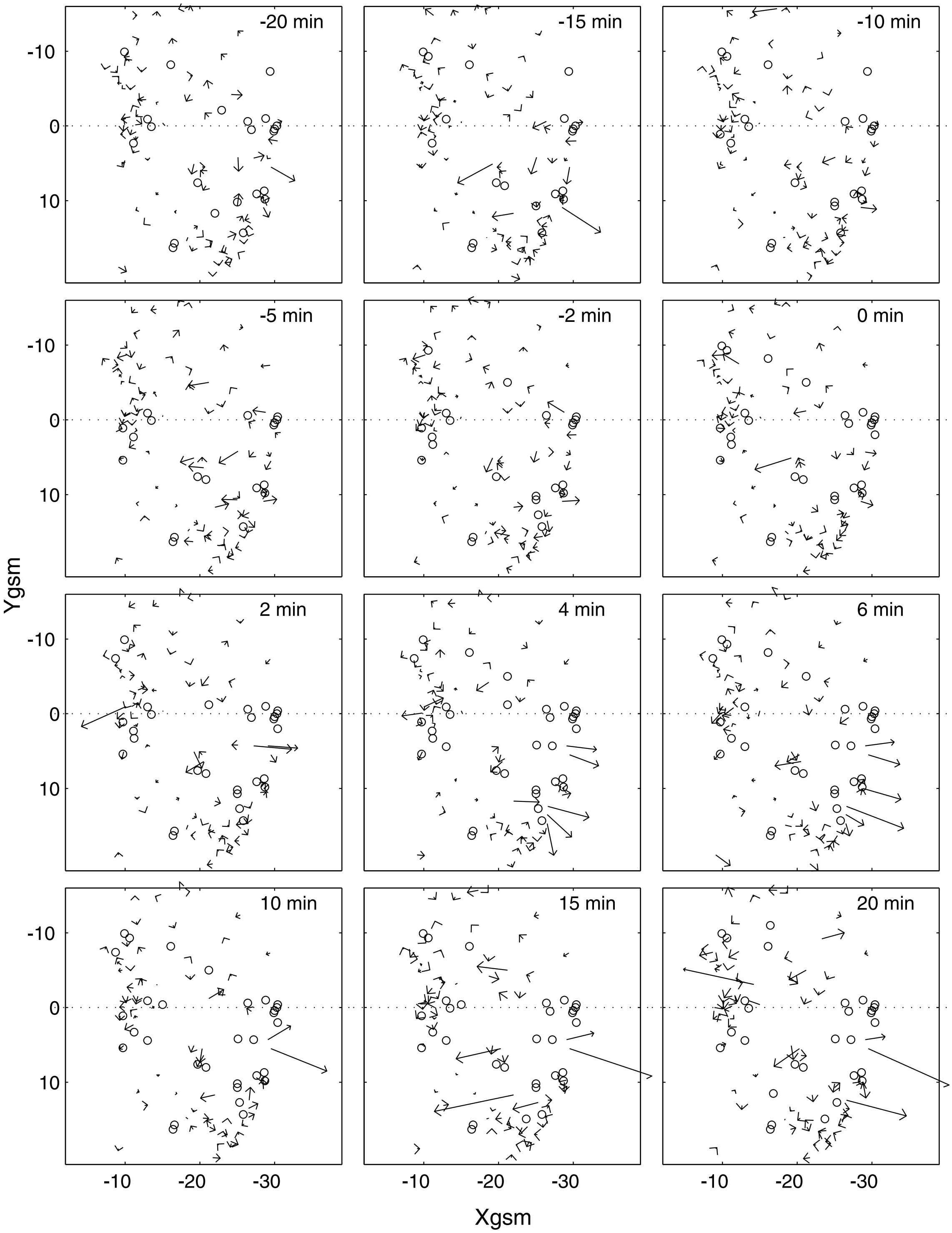
In reality, turbulent fluctuations in space plasmas are generally composed of the simultaneous coexistence of propagating modes and intermittent nonlinearly interacting spatiotemporal structures [32,33]. The "physics" of the bimodal state of such type of admixture of turbulent fluctuations may be understood from the point of view of the development and interactions of coherent structures arising from plasma resonance sites and the ensuring dynamical complexity resulting from such developments and interactions. In this article, we shall consider the dynamical complexity in space plasmas from such a concept. Sample results of direct numerical simulations and dynamical modeling including the calculated fluctuation probability distribution functions and local intermittency measures based on the wavelet transforms are provided to characterize the sporadic, localized, and scale-dependent nature of the intermittent turbulence. The concepts of multifractals, scale invariance and symmetry-breaking will be introduced. Observational examples demonstrating the existence of such phenomena and the associated invariance properties of intermittent space plasma turbulence will be provided.

Applications of the ideas of dynamical complexity to space plasmas were relatively recent. The first contact of these ideas with space plasmas was probably contained in a series of papers addressing the apparent low-dimensional chaotic behavior related to the dynamics of the Earth's magnetosphere [4,65,101,103,104,105]. These ideas were followed by a paper by Chang [22] that suggested the possibility of interpreting such phenomenon from the point of view of forced and/or self-organized criticality (FSOC) surmised from the method of the dynamic renormalization group (DRG). This suggestion was moti-



Space Plasmas, Dynamical Complexity in, Figure 1
Vector representation of the components of the magnetic field (B_y, B_z), ion plasma flow (V_y, V_z), and current density (J_x, J_z) during a magnetic substorm (i. e., when there was intense auroral activity) on August 22, 2001 observed by the Cluster spacecraft at a downstream distance of about 19 Re (Earth radii) in the magnetotail showing the large variability of these plasma parameters. Reprinted from [82] with thanks to the European Geophysical Union

vated by the low-dimensional chaos papers as well as the interpretation of the observed large variability of plasma fluctuations in the Earth's magnetotail as the result of sporadic and localized current disruptions [78], Figs. 1 and 2. At approximately the same time, Lu and Hamilton [74] used the idea of self-organized criticality (SOC) of Bak et al. [3] to explain the curious scale-independent behavior of the solar flare occurrence rate on the flare sizes. Such ideas suggest that complex dynamical systems generally organize themselves into states with statistical properties describable by power laws. (A succinct tutorial review of the SOC concept may be found in Jensen [63]). An alternative viewpoint of dynamical complexity is based on the concept of multifractals [54,98] which is a generaliza-



Space Plasmas, Dynamical Complexity in, Figure 2

A temporal sequence of synoptic patterns of total plasma flow showing large variabilities constructed from superposed epoch analysis of 102 magnetic substorm events from Geotail spacecraft observations. The scale is 100 km/s per Re (Earth radius). The acronym GSM stands for Geocentric Solar Magnetospheric coordinate system. The *X*-axis points to the Sun. The *XZ*-plane contains the Earth's dipole axis, and the *Y*-axis completes the right-handed coordinate system in units of Re. Reprinted from [80] with thanks to the American Geophysical Union

tion of the original ideas of fractals introduced by Mandelbrot [83]. Burlaga [12,13,14] was the first to incorporate such ideas to interpret the intermittent turbulent behavior of the solar wind. Other solar wind studies based on such a point of view followed, e. g. [16,17,18,19,20,58, 59,60,84,102]. These ideas were followed by Consolini et al. [41] and Consolini [37] in their interpretations of the observed fractal time series of the Auroral Electrojet (AE) indices.

Since then, a flurry of activities has blossomed into the current investigations of the phenomenon of complexity in space plasmas. Perhaps the most noteworthy are the connection of such phenomena with the traditional ideas of intermittent turbulence in plasmas and the development of the concept of crossovers (transitions) from one criticality to another and onto multifractals [27,28]. This article will provide simple descriptions of some of these modern ideas in conjunction with actual observational results and numerical simulations.

Due to the limitation of space, this entire article will center on the discussion of the complexity of space plasmas from the fluid description. It should be understood that situations also exist in space plasmas where the physical phenomena are related to the complexity of kinetic plasma effects, see, e. g., [33,111,115]. Since the purpose of this article is to introduce to the readers the basic concepts of – and recent developments in – the physics of dynamical complexity in space plasmas, only those primary references that touch upon these introductory ideas are included. No attempt has been made to include the hundreds of outstanding contributions that are available on the subject.

The contents of this article are structured as follows. In the next section, the concepts of plasma resonances and coherent structures are introduced. This is followed by a discourse on coarse-grained dissipation and magnetic reconfiguration. The ideas of non-Gaussian probability distributions of turbulent fluctuations, wavelet transforms and intermittency are then discussed in the two subsequent sections. Next, we introduce the basic ideas of multifractals in terms of structure functions, partition functions, singularity measures and rank-ordered multifractal spectra in Sect. "Multifractals". The concepts of scale invariants, forced and/or self-organized criticality are described in Sect. "Invariant Scaling" along with a discourse relating these ideas to multifractals. In Sect. "Dynamical Modeling – The Lu–Klimas Magnetic Field Reversal Model", we discuss the utility of dynamical models. A simple one-dimensional model, the Lu–Klimas model, is provided to illustrate the usefulness of such methods. This is followed by a brief discussion of what future details in the study and research of dynamical phenomena of complexity related to space plasmas. In addition to the primary list of references that are cited in the article, we also provide a secondary list of review articles and books in the bibliography for further in-depth study.

Plasma Resonances and Coherent Structures

Plasmas are known for their ability to form numerous types of coherent structures of varied sizes with scales that are generally much larger than those of the constituent particles (ions, electrons, neutrals) of the plasma medium. The reason behind this is closely related to the nonlinear and long range forces of the dynamical couplings among the constituent particles and species interacting within the electromagnetic environment.

Examples of coherent structures abound in the literature on theories and observations of nonlinear space plasma processes. They may appear as convective forms, nonlinear solitary structures, pseudo-equilibrium configurations, or other types of spatiotemporal entities. They may be locally generated or convected from elsewhere. Some of these structures may be more stable and long-lived; others may be less stable and relatively short-lived and sometimes partially formed. Generally, such structures are not purely laminar entities; they are likely composed of bundled fluctuations of all frequencies and scales. Due to the nature of the physics of complexity, it would be futile to attempt to evaluate and/or study the details and stabilities of each of these countless interacting structures or partially formed structures, though some basic understanding of the different types of structures would be helpful in the comprehension of the full stochastic complexity that results from their underlying nonlinear dynamics.

As an example of such coherent structures, let us consider the behavior of Alfvénic flux tubes in magnetized plasmas. We shall base our theoretical discussion by considering the plasma medium as a single, charge-neutral, conducting fluid moving in and interacting self-consistently with its electromagnetic environment. Because the plasmas in the space environment are generally highly rarefied, we can typically neglect the dissipative effects from particle collisions and assume the plasma is perfectly conducting and inviscid. The most elemental mathematical formulation for magnetized plasmas under such simplifying assumptions is the so-called ideal incompressible magnetohydrodynamics (MHD). The basic equations are the equations of continuity and motion, and the relevant Maxwell's equations where the displacement current is neglected for nonrelativistic motion.

$$\nabla \cdot \mathbf{v} = 0 \quad \text{(continuity equation for incompressible medium)} \tag{1}$$

$$\rho \mathrm{d}\mathbf{v}/\mathrm{d}t = \mathbf{j} \times \mathbf{B} - \nabla p \quad \text{(equation of motion for an inviscid medium)} \tag{2}$$

$$\nabla \cdot \mathbf{B} = 0 \text{ (Gauss' law)} \tag{3}$$

$$\partial \mathbf{B}/\partial t + \nabla \times \mathbf{E} = 0 \quad \text{(Faraday's law)} \tag{4}$$

$$\mathbf{j} = \nabla \times \mathbf{B} \quad \text{(Ampère's law neglecting the displacement current)} \tag{5}$$

where $\mathrm{d}/\mathrm{d}t = \partial/\partial t + \mathbf{v} \cdot \nabla$ is the convective derivative, ρ is the density of the medium, $\mathbf{v}$ is the velocity, $\mathbf{j}$ is the current density, $\mathbf{B}$ is the magnetic field, $\mathbf{j} \times \mathbf{B}$ is the Lorentz force, p is the fluid pressure, and $\mathbf{E}$ is the electric field. These equations are written in SI units with the vacuum permeability set equal to 1.

To complete the set of equations of such a formulation, we complement the above equations with the Ohm's law for a perfectly conducting medium:

$$\mathbf{E} + \mathbf{v} \times \mathbf{B} = 0 \tag{6}$$

These equations may be combined to form the following set of equations of motion and induction:

$$\rho \mathrm{d}\mathbf{v}/\mathrm{d}t = (\mathbf{B} \cdot \nabla)\mathbf{B} + \cdots \tag{7}$$

$$\mathrm{d}\mathbf{B}/\mathrm{d}t = (\mathbf{B} \cdot \nabla)\mathbf{v} \tag{8}$$

where the eclipses represent the gradient of the total pressure $(p + B^2/2)$. Standard arguments lead Eqs. (7) and (8) to a linearized wave equation characterizing the fundamental propagation of small fluctuations of $\mathbf{v}$ and $\mathbf{B}$. This wave equation admits the well-known Alfvén waves which are transverse waves to the magnetic field $\mathbf{B}$ and propagate along the magnetic field direction. The phase velocity v_A of the Alfvén wave is $B/\sqrt{\rho}$, i. e., proportional to the strength of the magnetic field and inversely proportional to the square root of the density ρ of the plasma. For such waves to propagate, the operators on the right hand sides of (7) and (8) must not vanish, i. e., $\mathbf{B} \cdot \nabla \rightarrow i\mathbf{k} \cdot \mathbf{B} = k_\parallel \neq 0$. Therefore, the propagation vector $\mathbf{k}$ must contain a field-aligned component $k_\parallel$. When the parallel component $k_\parallel$ of the propagation vector vanishes (i. e., at the resonance sites), the fluctuations are localized. Around these resonance sites (usually in the form of three-dimensional space curves), it may be shown that the fluctuations – which will try to propagate away as Alfvén waves – are held back by the background magnetic field and the plasmas medium, thereby forming the so-called Alfvénic coherent structures (which are actually domains of bundled stochastic fluctuations).

Coarse-Grained Helicity

We now consider such magnetized domains near the Alfvenic resonance sites. For an ideal MHD system, any physically acceptable magnetic field must satisfy the solenoidal condition (Gauss' law): $\nabla \cdot \mathbf{B} = 0$. Also, any variation of the field away from the initial value must satisfy the constraints (4) and (6), i. e., Faraday's law and Ohm's law for infinite conductivity.

It may be easily demonstrated that these constraints are equivalent to an infinite set of integral constraints involving the helicity K, such that

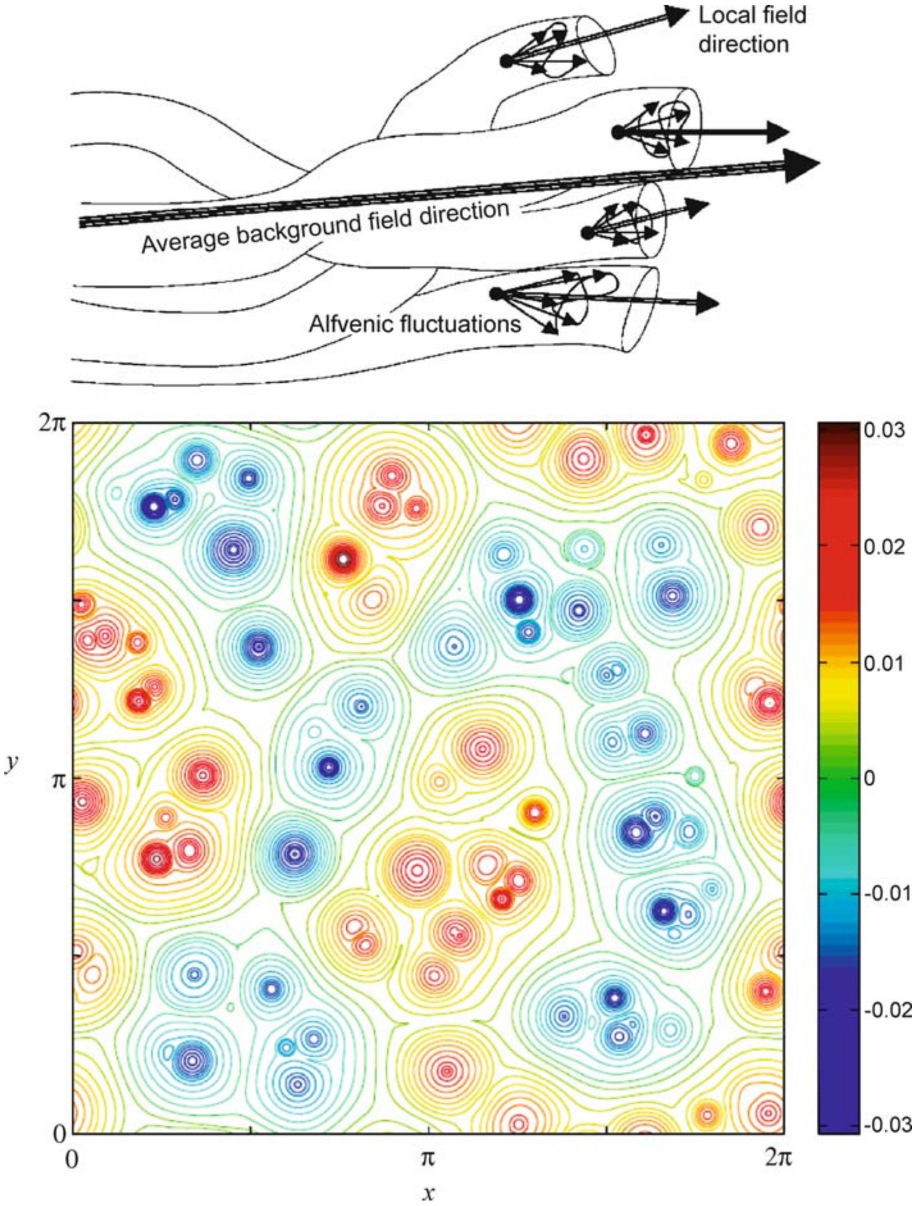
$$K = \int_V \mathbf{A} \cdot \mathbf{B} \mathrm{d}V \tag{9}$$

is an invariant for any single connected volume V enclosed by a flux surface, where $\mathbf{A} = \nabla \times \mathbf{B}$ is the vector potential.

Taylor's Conjecture

We are interested in the situation where there are domains with stochastic turbulent fluctuations within which field lines merge and mix indistinguishably. Thus, it will be difficult – in fact, unpractical – to discuss the topology of individual field lines. Nonetheless, it was suggested by Taylor [114] that when the volume integral of (9) is taken over the stochastic region, the coarse-grain averaged helicity in a relaxing state would be essentially conserved. As the domain of the stochastic region near the Alfvenic resonance site relaxes to a statistically stationary minimum energy state under the constraint of this conjecture of conservation of coarse-grained helicity, it may be shown using the variational principle that such a domain (coherent structure) will be essentially force-free in the sense that $\mathbf{j} \times \mathbf{B} = 0$ where $\mathbf{j}$ and $\mathbf{B}$ are the mean current density and magnetic field, respectively.

We are, of course, interested in dynamical states that are far from equilibrium. Thus, in visualizing the relaxed states from the point of view of this *Taylor's conjecture*, we shall consider timescales such that "nearly coherent" structures are formed. These structures actually move, mix



Space Plasmas, Dynamical Complexity in, Figure 3

***Top*: Schematics of tangled flux tubes in the solar wind. Each flux tube is characterized by a local magnetic field direction aligned approximately with the background field and the presence of Alfvénic fluctuations makes the magnetic field vector wander randomly about this direction. Moving across the tubes, strong intermittent interactions of predominantly nonpropagating (i. e., resonant) fluctuations are expected. Reprinted from [10] with permission from Elsevier. *Bottom*: 2D MHD simulation of cross-sectional view of interacting flux tubes for homogeneous intermittent turbulence. *Colors* represent current intensities and directions. Adapted from [33] with thanks to the American Institute of Physics**

and sometimes merge together while immersed in an otherwise turbulently diffusing plasma medium.

To obtain some physical insight of the geometries of these magnetic coherent structures, let us consider the special situation in the solar wind and make the reasonable assumption that the perturbed magnetic field fluctuations are much smaller than – and essentially transverse to – the mean magnetic field B_0 which will be temporarily assumed to be uniform for the current discussion [89]. Thus, let us write $\mathbf{B} = (\delta B_x, \delta B_y, B_0)$, where z is in the direction of the mean magnetic field, and (x, y) are orthogonal coordinates normal to z. The force-free condition for constant B_0 and $\nabla \cdot \mathbf{j} = 0$ then leads approximately to the scalar condition $\mathbf{B} \cdot \nabla j_z = 0$, obtained by taking the z-component of the curl of $\mathbf{j} \times \mathbf{B} = 0$. We have, then,

$$B_0 \partial j_z / \partial z = -(\delta B_x \partial / \partial x + \delta B_y \partial / \partial y) j_z \, . \tag{10}$$

For convenience, let us introduce the flux function ψ by writing $(\partial \psi / \partial y, -\partial \psi / \partial x) = (\delta B_x, \delta B_y)$ for the perturbed transverse components of the magnetic field in the (x, y) directions such that the Gauss law of magnetism, $\nabla \cdot \mathbf{B} = 0$ is satisfied. Then, j_z and ψ are governed by Eq. (10) and the Ampere's law, $\nabla \times \mathbf{B} = \mathbf{j}$.

A simple example of the flux function and axial current density satisfying the above conditions would be the class of circularly cylindrical solutions of $\psi(r)$ and $j_z(r)$, [25,26]. Generally, the solutions would be more involved because of the variabilities of the local conditions of the plasma and the three-dimensional geometry. Moreover, the dynamic coherent structures with the inclusion of plasma pressure and other modifying effects (including electron-inertia terms) would be even more complicated. However, we expect these structures to be in the form of nearly field-aligned flux tubes, Fig. 3 [10,25,26, 136,137,138]. Existence of Alfvénic flux tubes in the magnetopause and magnetotail have also been suggested by Tetreault [116,117] and Chang [23,24], respectively.

Generally, there exist various types of propagation modes (whistler modes, electromagnetic ion cyclotron waves, etc.) in magnetized plasmas. Thus, we envision a corresponding number of different types of plasma resonances and associated coherent structures that typically characterize the dynamics of the plasma medium under the influence of an electromagnetic background.

Coarse-Grained Dissipation and Magnetic Reconfiguration

When coherent Alfvénic flux tubes with the same polarity migrate toward each other, strong local magnetic shears are created, Fig. 4. Wu and Chang [136,137,138] have demonstrated that the existing sporadic nonpropagating fluctuations in the strong local shear region, particularly those close to the neutral sheet (i. e., at the location where the local magnetic field vanishes), will stay in the region and continue to interact nonlinearly. On the other hand, fluctuations away from the neutral sheet region, are nonresonant and will therefore propagate away along magnetic field lines as Alfvén waves. Combined with the magnetic shear geometry, the resonant fluctuations will induce a nonlinear instability near the neutral sheet region, which will produce more fluctuations – nonresonant and resonant. The nonresonant fluctuations will again propagate away as Alfvén waves while the resonant ones will join the other resonant fluctuations and interact nonlinearly, thereby broadening the resonance region, Fig. 5.

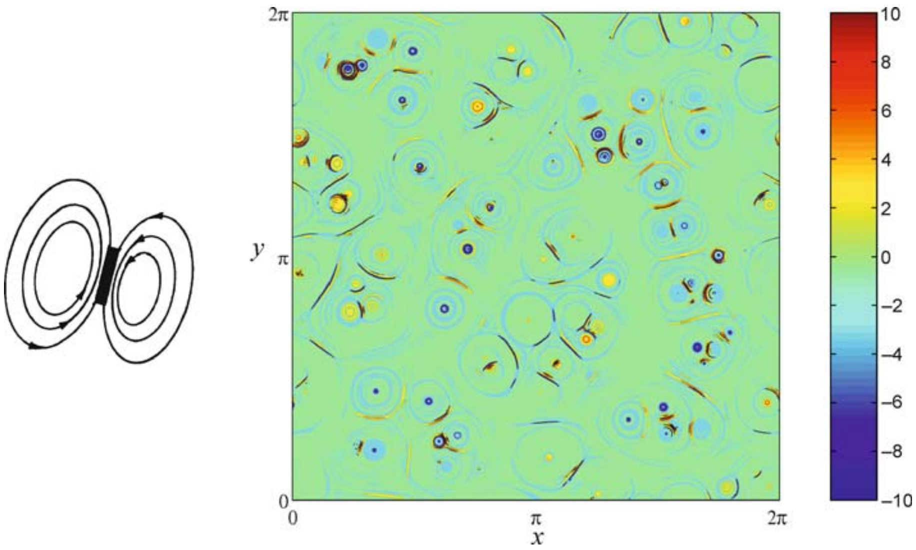
This combined phenomenon of "coarse-grained dissipation" depletes the energy originally contained in the coarse-grained magnetic fields near the shear region, initiating a reconfiguration of the coherent structures. In coarse-grained sense, it breaks some of the closed field lines of each of the coherent structures and then reconnects them into single closed field lines. And this process continues with the system adjusting intermittently with the surrounding environment until all free energies are exhausted, eventually leading to the formation of one single combined coherent structure with one set of coarse-grained concentric closed field lines, Fig. 6.

The final state of the resulting coherent structure will have less average energy due to the combined dissipation of Alfvénic propagation of nonresonant fluctuations and nonlinear interactions of the resonant fluctuations (i. e., resonance broadening). Such is the manifestation of continuous magnetic topological reconfiguration due to the dynamic "fluctuation-induced nonlinear instability". And this merging process may repeat over and over again among the coherent structures of the same polarity, from the largest scales to the smallest scales where kinetic effects may have to be included.

On the other hand, when coherent structures of opposite polarities approach each other due to the forcing of the surrounding plasma, they might repel each other, scatter, or induce magnetically quiescent localized regions.

Under any of the conditions of the above interaction scenarios, new fluctuations will be generated. And these new fluctuations can provide new resonance sites, thereby nucleating new coherent structures of varied sizes. This kind of tangled geometry of interacting flux tubes leading to dynamical complexity has been described by Bruno et al. [10] who first deduced their existence in the solar wind, as "cooked spaghetti", Fig. 3 (top panel), and demonstrated by Wu and Chang [136,137,138], Fig. 3 (bottom panel), Matthaeus et al. [90], Fig. 7, and others with MHD simulations. And they are the source of the observed sporadic and localized current disruptions in the Earth's magnetotail and elsewhere [75,78,82].

Topological reconfigurations of such nature occur quite frequently for the dynamical interactions of coherent structures in space plasmas and are not limited just to flux tubes; for example, Sundkvist et al. [109] and Alexandrova et al. [1] have observed intermittent interactions of drift Alfvén vortices in the cusp and magnetosheath, respec-



Space Plasmas, Dynamical Complexity in, Figure 4
Cross-sectional views of interacting Alfvénic flux tubes. *Left*: Schematic of merging. Arrows indicate directions of magnetic field and blackened area indicates location of strong shear. *Right*: 2D MHD simulation of the current intensities for homogeneous intermittent turbulence. *Colors* represent the intensities and directions. Adapted from [137] with permission from Elsevier

tively. Similarly, such enhanced intermittency at the intersection regions of whistler coherent structures has also been surmised by Consolini and Lui [40] and Consolini et al. [42] in the plasma sheet. All such intermittent interactions are quite akin to the avalanche phenomenon prevalent in sandpile models [3,35,36,62,133]. They are the origin of the various observed magnetic reconnection signatures in space plasmas.

This stochastic behavior of the interactions of the plasma coherent structures is a phenomenon of "dynamical complexity" as defined in Sect. "Definition of the Subject", which, for emphasis, is reproduced below:

"Dynamical complexity" is a phenomenon exhibited by a nonlinearly interacting dynamical system within which multitudes of different sizes of large scale "coherent structures" are formed, resulting in a complicated global nonlinear stochastic behavior for the dynamical system – vastly different from that could be surmised from the original dynamical equations.

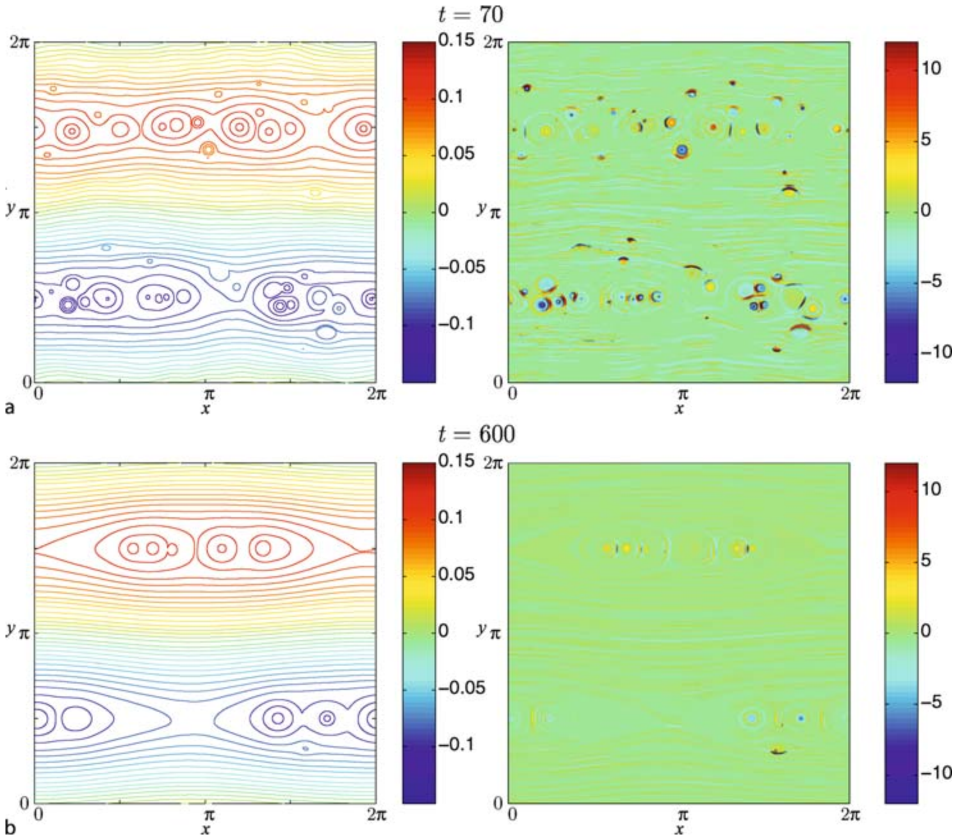
Non-Gaussian Probability Distribution Functions

The fluctuations that are induced by the interactions and mergings of coherent structures are sporadic and localized. Since the coherent structures are numerous and outsized, we expect the fluctuations within the interaction regions of these structures (resonance overlap regions) to be large and occur relatively more frequently than those that would have been expected from a medium of the original minute plasma particles (electrons, ions and neutrals). A technique useful in gauging the degree of such effects is by studying the shapes of the probability distribution functions (PDFs) of intermittent fluctuations at varying scales.

To demonstrate this, let us refer to one of the 2D MHD simulations described previously [32,33,136,137,138]. (In the following discussions, the measured fluctuations may be those of any physical property of the MHD medium. We shall choose the physical property as the strength of the magnetic field $B(x_i)$ to render the discussions more specific). Consider, for example, the spatial series of the strength of the magnetic field $B(x_i)$ for a given time t and at fixed y, with $x_i = i\delta$ where $i = 0, 1, 2, \ldots, N$, and δ is the grid size of the simulation. We can then form the coarse-grained differences

$$\delta B_i^2(\Delta) = B^2(x_i + \Delta) - B^2(x_i) \qquad (11)$$

within the interval with $\Delta = k\delta$ (k: an integer) and generate the probability distribution function $P(\delta B^2, \Delta)$. Figure 8 displays the calculated results of $P(\delta B^2, \Delta)$ for a 2D MHD simulation for the homogeneous case for several coarse-grained scales, Δ. From this figure, we note that the distribution for each scale falls nearly onto a smooth curve with the exception at the tails where scatterings from the mean are more visible. And the shapes of the PDF curves deviate more and more from that of a Gaussian at smaller and smaller scales.



Space Plasmas, Dynamical Complexity in, Figure 5
2D MHD simulation of coarse-grained dissipation in sheared magnetic field. *Left*: Contours of magnetic flux. *Right*: Corresponding current density distributions. Initially, new fluctuations are excited and the resonant fluctuations begin to interact nonlinearly near the neutral sheet region (near $y = \pi/2, 3\pi/2$). Eventually, the field lines are reconnected in the coarse-grained sense. Adapted from [137] with permission from Elsevier

Such PDFs sometimes satisfy the one-parameter scaling form:

$$P(\delta B^2, \Delta)\Delta^s = P_s(\delta B^2/\Delta^s) \tag{12}$$

where s is the parameter (or scaling exponent) such that all PDFs essentially collapse onto one master scaling function P_s. Such scaling behavior seems to be approximately satisfied for the above simulated result with $s \approx 0.335$ (Fig. 9), although closer examination reveals some discrepancies – particularly in the tail regions [33]. The more subtle nature of the scaling properties of such PDFs will be considered further in Sect. "Multifractals".

We shall now proceed to make contact with in-situ spacecraft observations. Until recently, most of the turbulence data in space came from measurements obtained by one single spacecraft. The data are then analyzed based on the Taylor hypothesis [113] which assumes that the transit time of eddies (bundled fluctuations) is much less than the characteristic evolution time. For fully developed MHD turbulence, the characteristic eddy turnover (evolution) time may be estimated:

$$t(\text{evolution}) \sim (\lambda/2\pi)/[(\delta B/B_0)v_A] \tag{13}$$

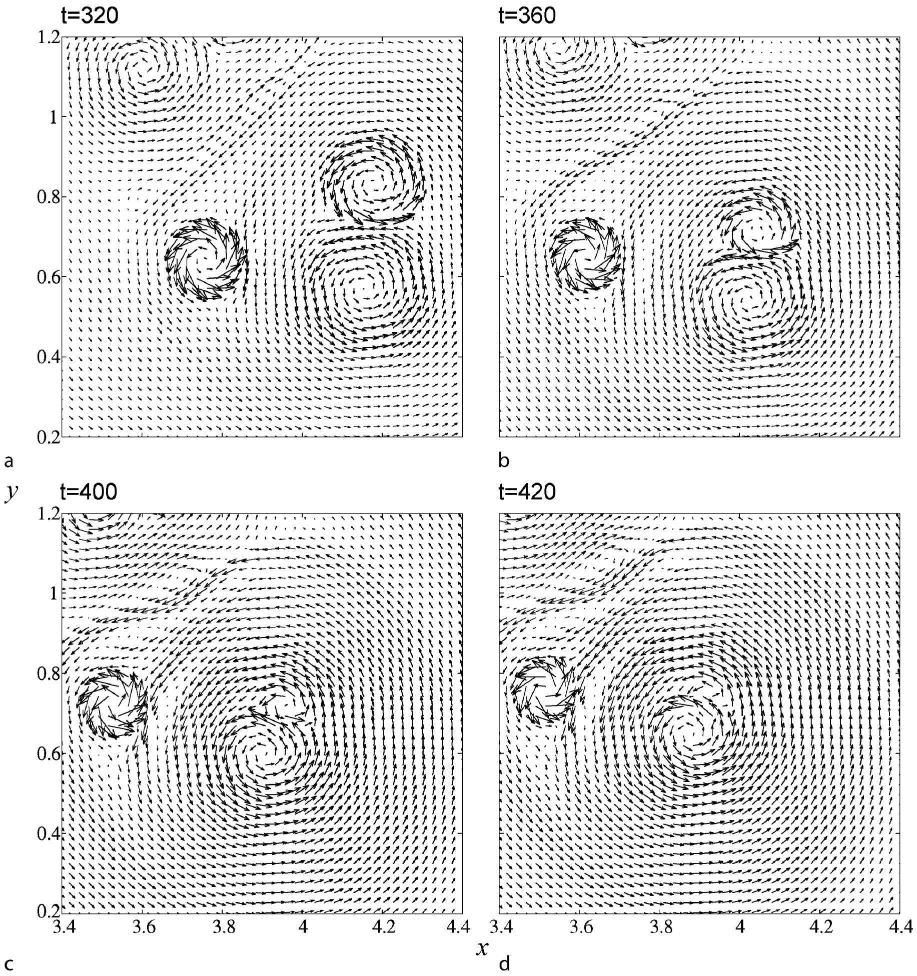
where λ is the typical size of the eddy, v_A is the Alfvén speed based on the mean magnetic field B_0 and δB is the average fluctuating magnetic field of the eddy in question. The transit time of the eddy is:

$$t(\text{transit}) \sim \lambda/v_s \tag{14}$$

where v_s is the relative streaming speed between the plasma medium and the spacecraft.

Thus, as it has been demonstrated by Matthaeus and Goldstein [88] that the requirement of $t(\text{evolution}) \gg t(\text{transit})$ becomes:

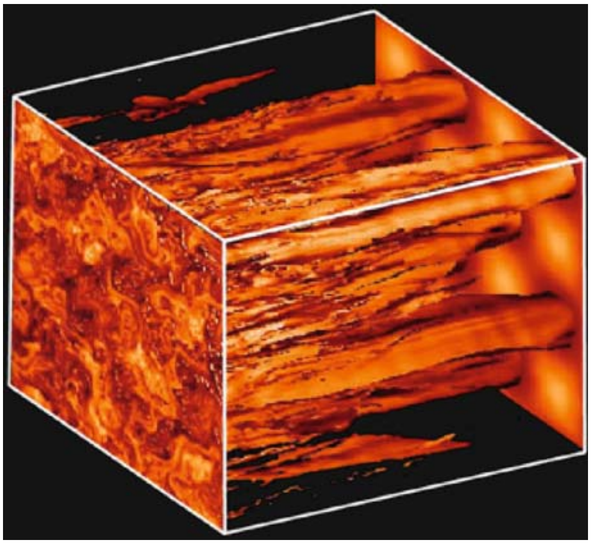
$$v_s/v_A \gg 2\pi(\delta B/B_0) \tag{15}$$



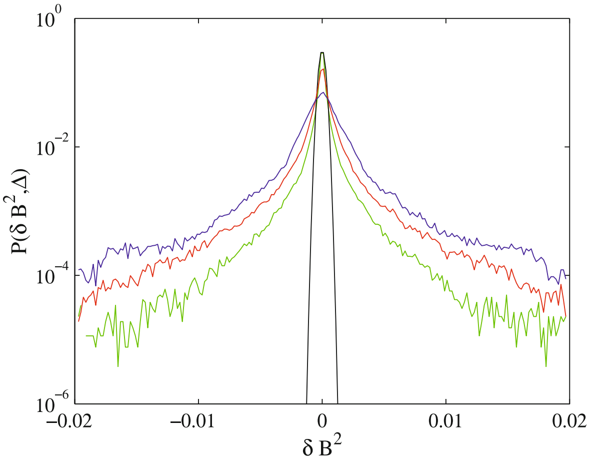
Space Plasmas, Dynamical Complexity in, Figure 6
2D MHD simulation of merging of flux tubes. *Arrows* indicate magnitudes and directions of magnetic fields. Reprinted from [137] with permission from Elsevier

As pointed out by Tu and Marsch [120], at 1 AU for solar wind observations the wind speed (which is essentially the relative speed between the plasma medium and the spacecraft since it is much larger than the spacecraft speed) is of the order of 10 times that of the ambient Alfvén speed and $\delta B/B_0 \simeq 0.5$. Therefore, condition (15) is satisfied for solar wind turbulence in that region. Because the spacecraft speed is much smaller than the solar wind speed, the data gathered by instruments aboard the spacecraft essentially sample the spatial fluctuations in the direction of the solar wind, particularly at times when the solar wind turbulence is homogeneous at large scales and fully developed. Thus, for example, the time series of such measurements of the strength of the magnetic field $B(t_i)$, where $t_i = i\delta$ with $i = 0, 1, 2, \ldots, N$ and δ is the sampling interval, may be interpreted based on the hypothesis as the spatial series of the ambient turbulence fluctuations with $x = v_s t$ (x: direction of the solar wind velocity). We can then construct the probability distribution function $P(\delta B, \tau)$ or $P(\delta B^2, \tau)$ with the coarse-grained scale $\tau = k\delta$ (k: integer) in analogy to the calculations related to the simulation results discussed above.

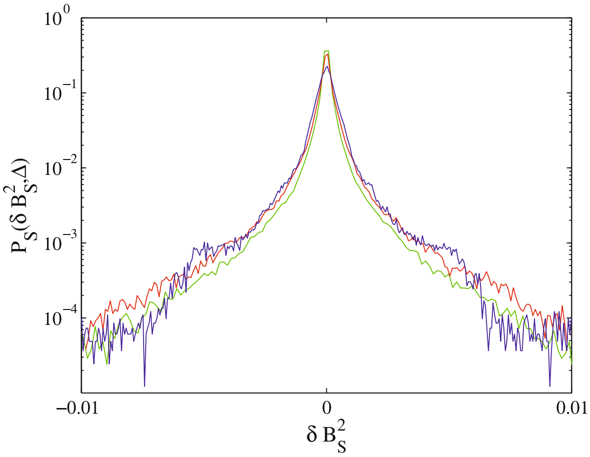
Non-Gaussian probability distribution functions from such statistics have been commonly observed in the solar wind for fluctuations related to the magnetic, velocity, density and other field variables, e. g. [15,47,48,56,71,72,85,107,108]. Figure 9 depicts the typical PDFs of magnetic field fluctuations obtained from such an analysis for both the slow ($\sim (400 \pm 200)$ km/s) and fast ($\sim (700 \pm 100)$ km/s) solar wind [108]. The similarities between the simulated results and results from such observations as exhibited in Fig. 9 are quite striking.



Space Plasmas, Dynamical Complexity in, Figure 7
MHD simulation of a two-component model, with 80% of its energy in two-dimensional modes, exhibiting tangled flux tubes with considerable transverse complexity. Reprinted from [90] with permission from the American Astronomical Society



Space Plasmas, Dynamical Complexity in, Figure 8
Probability distribution function $P(\delta B^2, \Delta)$ from a 2D MHD simulation for homogeneous turbulence for $\Delta = 2$ (*green*), 8 (*red*), 32 (*blue*) units of grid spacing. The *black* curve is the PDF for Gaussian fluctuations. Reprinted from [34] with permission from Springer Science and Business Media



Space Plasmas, Dynamical Complexity in, Figure 9
Scaled probability distribution function $P_s(\delta B_s^2, \Delta)$ with $\delta B_s^2 = \delta B^2/\Delta^s$, $s = 0.335$, and $\Delta = 2$ (*green*), 8 (*red*), 32 (*blue*) units of grid spacing. Adapted from [33] with thanks to the American Institute of Physics

One popular empirical technique in expressing the shape of the PDFs of the observed intermittent turbulent fluctuations is the Castaing distribution. In 1990, Castaing et al. [21] suggested that the intermittent fluctuations might perhaps be viewed as an ensemble of subsets of fluctuations ξ, each subset having a normal distribution:

$$P_\sigma(\xi) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\xi^2/2\sigma^2) \tag{16}$$

where σ is the variance. The intermittency is then assumed to be due to the fluctuations of the variances that satisfy a log-normal distribution:

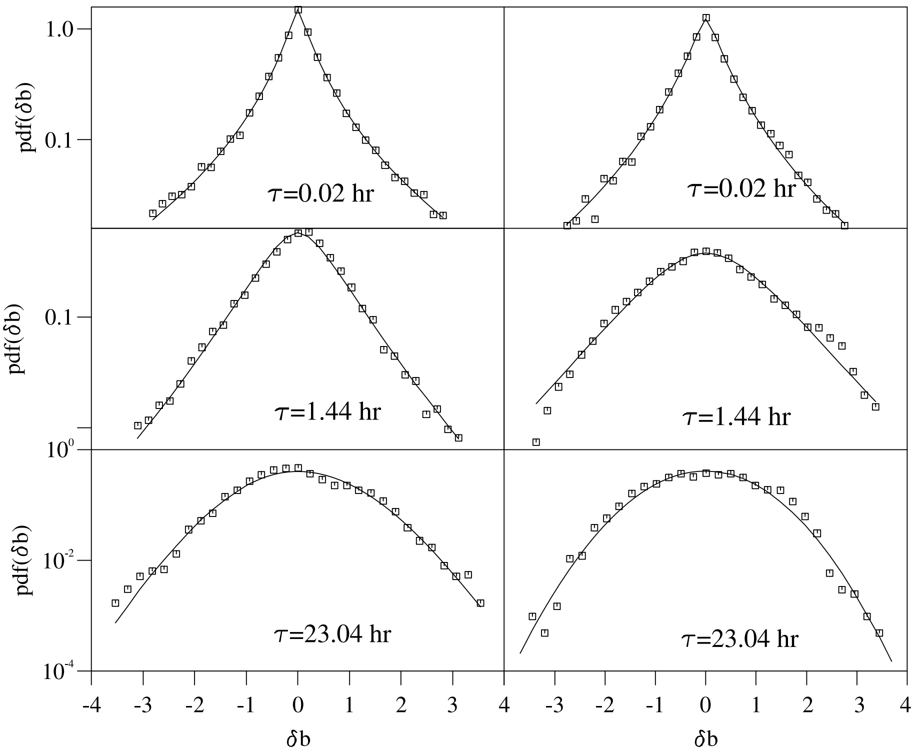
$$Q_\lambda(\sigma)\mathrm{d}\sigma = \frac{1}{\lambda\sqrt{2\pi}} \exp\left(-\frac{\ell n^2(\sigma/\sigma_0)}{2\lambda^2}\right) \mathrm{d}\ell n\sigma \tag{17}$$

where σ_0 is the most probable variance of ξ, and λ is the variance of $\ell n\sigma$. Combining (16) and (17) gives the Castaing distribution

$$\Pi_\lambda(\xi) = \frac{1}{2\pi\lambda}\int_0^{+\infty} \exp\left(-\frac{\xi^2}{2\sigma^2}\right) \exp\left(-\frac{\ell n^2(\sigma/\sigma_0)}{2\lambda^2}\right) \frac{\mathrm{d}\sigma}{\sigma^2}\,. \tag{18}$$

Such fits of data to (18) are shown as solid curves in Fig. 10. How well does such an ansatz characterize the data depends very much on the type and scale of the fluctuations [47]. Other empirical fits such as the kappa-distribution and gamma-distribution have also been suggested, but we do not have space here to enumerate and discuss these and other phenomenological models.

Similar observational results have been detected for the magnetic field and velocity fluctuations in the plasma sheet [2,134], in the cusp [44,139], for the electric field



Space Plasmas, Dynamical Complexity in, Figure 10
PDF for δB calculated from the observational data in the (*left*) fast solar wind (∼ 700 km/s) and (*right*) slow solar wind (∼ 400 km/s) streams. *Solid lines* represent fits obtained with the model suggested by Castaing et al. [21]. Reprinted from [108] with permission from Elsevier

fluctuations in the auroral zone [111,112], as well as for fluctuations of other field variables, though caution must be exercised in interpreting some of these results as the Taylor hypothesis may or may not be satisfied in the observational regions.

Wavelet Transforms and Intermittency

Traditional procedure for analyzing fluctuations is to perform Fourier transforms of the observational, numerical or experimental data. The basis functions of the Fourier transform are sines and cosines. Each such basis function extends in space (or time) all the way from minus infinity to plus infinity. Thus, Fourier transform essentially wipes out the localized information of the intermittent fluctuations resulting from the confined interactions of the coherent structures. Therefore, for intermittent turbulence, it will be more appropriate to consider transforms whose basis functions are localized and scale dependent instead of sines and cosines. Such a transform has the generic name of "wavelet transform" [43,46]. Below, we restrict our discussions to one-dimensional wavelet transforms.

To be more specific, a wavelet transform replaces (transforms) the fluctuating spatial (or temporal) series into the coefficients of a set of basis functions that are localized in space (or time) with different scales in space (or time). The basis functions are usually chosen such that they are square-integrable and generated by a single mother wavelet, a rescalable and translatable function $\psi[(x-b)/a]$ where a is the scale and b defines the translation along x. Two of the more popular mother wavelets are the complex-valued Morlet wavelet [94] defined by

$$\psi_0(y) = \mathrm{e}^{ik\cdot y}\mathrm{e}^{-|y|^2/2} \tag{19}$$

and the Haar wavelet [52] given by

$$\psi_0(y) = \begin{cases} 1 & \text{if } y \in [0, 1/2); \\ -1 & \text{if } y \in [1/2, 1),\ \text{and} \\ 0 & \text{if } y \notin [0, 1)\,. \end{cases} \tag{20}$$

To generate the full set of functions, we simply replace the argument in (19) or (20) by $(x-b)/a$ and allow the parameters (a, b) to vary within the domain of interest.

The Morlet wavelet is appealing to the physicists because it fits in with the usual understanding of wave packets. And due to its continuous nature, the transform is amenable to analytical manipulations. One disadvantage of the Morlet wavelets is that they are not entirely orthogonal to each other. The Haar wavelet, on the other hand, is truly local and the basis functions are real and orthogonal to each other. The simple and discrete nature of the Haar wavelets renders it ideal for numerical applications. One disadvantage of the Haar wavelets is that they are not symmetrical with respect to the midpoint, thus producing slanted spectral plots. This asymmetry, however, becomes less and less noticeable for smaller and smaller scales.

To obtain the coefficients of a wavelet transform for a particular fluctuating spatial (or temporal) series $S(x)$, we evaluate the convolution integral as follows:

$$C(a,b) = (K/a)\int S(x')\psi[(x'-b)/a]\mathrm{d}x' \qquad (21)$$

where K is a normalization constant.

From $C(a,b)$, we can calculate, for example, the normalized power $\mathrm{NP}(a,b)$:

$$\mathrm{NP}(a,b) = |C(a,b)|^2 / \langle |C(a,b)^2| \rangle_x \qquad (22)$$

where $\langle \dots \rangle_x$ denotes the spatial (or temporal) average; and, here, the average is essentially the Fourier transformed power for the entirety of the fluctuations. In usual applications, only a truncated finite set of coefficients are used to perform practical calculations. The normalized power is a power spectrum of both scale and location; thus, it is sometimes called the Local Intermittency Measure, LIM. By definition, LIM is equal to 1 for the Fourier transform. Figure 11(top) is a color representation of the Local Intermittency Measure or LIM of the magnetic field intensity using the Haar transform for the 2D homogeneous MHD simulation discussed previously at a given time t and for fixed y. The scale-dependent, sporadic and localized powers of the fluctuations are vividly displayed. We note that the intermittency increases as the scale is reduced. And, the intensity can be rather strong at small scales – even when the fluctuations are strongly intermittent and localized.

Normalized power of wavelet transforms have been used to evaluate the Local Intermittency Measure LIM in the solar wind (e. g. [9,10]), for the AE index [38,39], in the magnetotail (e. g. [42,77,131]), in the cusp (e. g. [44, 139]), and in the auroral zone [112]. Figure 11 (bottom) is a color representation of the Local Intermittency Measure LIM using the Haar transform for the electrical field fluctuations in the transverse direction to the magnetic field in the auroral zone [112]. The characteristics of strong intermittency at small scales are clearly visible.

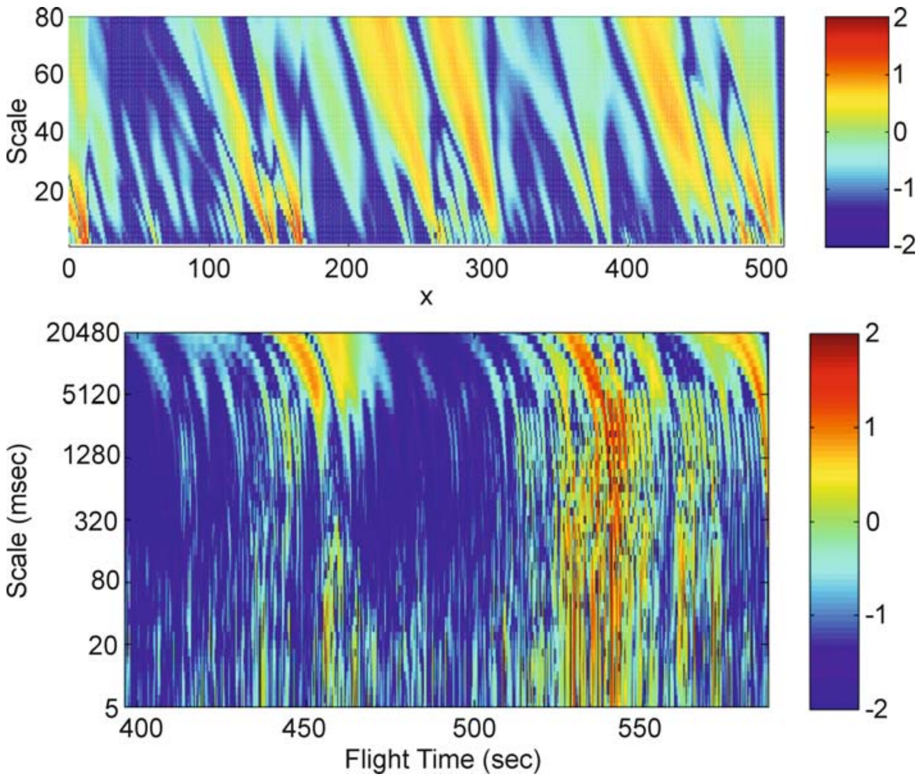
As it has been pointed out in Sect. "Non-Gaussian Probability Distribution Functions", intermittent fluctuations have non-Gaussian probability distributions especially at small scales. We may look for a quantity which provides a measure of the non-Gaussian nature of the fluctuations as a function of scale. A normalized scale dependent measure related to the fourth order moment of the coarse-grained probability distribution is called the Flatness. It indicates whether the data are more peaked or flatter relative to the Gaussian distribution. It is also called the Kurtosis. It has been suggested by Meneveau [91] that the average over space (or time) of the square of the normalized power is essentially the scale dependent Flatness of the probability distribution.

$$\text{Flatness} = \left\langle [\mathrm{NP}(a.b)]^2 \right\rangle_x \qquad (23)$$

For Gaussian probability distributions, the value of the Flatness may be shown to be equal to 3. For an intermittent event, the Flatness increases without bound at small scales. The rate at which the Flatness increases with the decrease of scale is a measure of intermittency. Figure 12(top) plots the values of the Flatness calculated as a function of scale corresponding to the fluctuations discussed above for the magnetic field fluctuations for the 2D MHD simulation for given y and fixed t. We note that the value of the Flatness becomes larger than 3 and continues to increase as the scale becomes smaller and smaller.

Flatness has been evaluated for the solar wind turbulence by Bruno et al. [10,11], Fig. 12(bottom). It is one of the tools used by Bruno et al. [10] to arrive at the conclusion that solar wind turbulence is the manifestation of interacting tangled flux tubes or "cooked spaghetti" as depicted in Fig. 3(top). Wavelet transforms have been used by various authors to evaluate Flatnesses of intermittent turbulence in other regions of space plasmas, e. g. [44,77, 112,131,139]. Other more generalized techniques of gauging the intermittency in space plasma turbulence have also been employed by Vörös [129] and Vörös et al. [130,132], which we shall not discuss here.

The normalized power and Flatness are only two tools that provide gauges of intermittency of a fluctuating series. One might ask: Are there other measures such that in their totality fully characterize the intermittency nature of the fluctuating series? This will be addressed briefly in the next section.



Space Plasmas, Dynamical Complexity in, Figure 11

***Top*: Normalized power or LIM for the intermittent magnetic field fluctuations *B*(*x*) using the Haar transform for a 2D homogeneous MHD simulation at given time *t* and for fixed *y*, in units of grid spacing. *Color* indicates intensity. *Bottom*: Normalized power or LIM for the intermittent transverse electric field fluctuations in the auroral zone. *Color* indicates intensity. Adapted from [112] with thanks to the American Geophysical Union**

Multifractals

Another popular modus operandi designed to study the phenomenon of intermittency is based on the concept of multifractals. As we have seen above, turbulence in space plasmas generally encompass fluctuations of all varieties and sizes, which interact and propagate throughout the plasma medium. For illustrative purposes, let us visualize some particular fluctuations that have conventional geometrical properties in a three-dimensional Euclidean space. Because of their sporadic and localized nature, it is easy to imagine that they generally cannot fill the full three-dimensional space that they occupy at a given time. Or said in another way, the space these fluctuations occupy is only a fraction of the full three-dimensional space. Such geometrical property was popularized by Mandelbrot [83] when he coined the word "fractals" or fractal geometry.

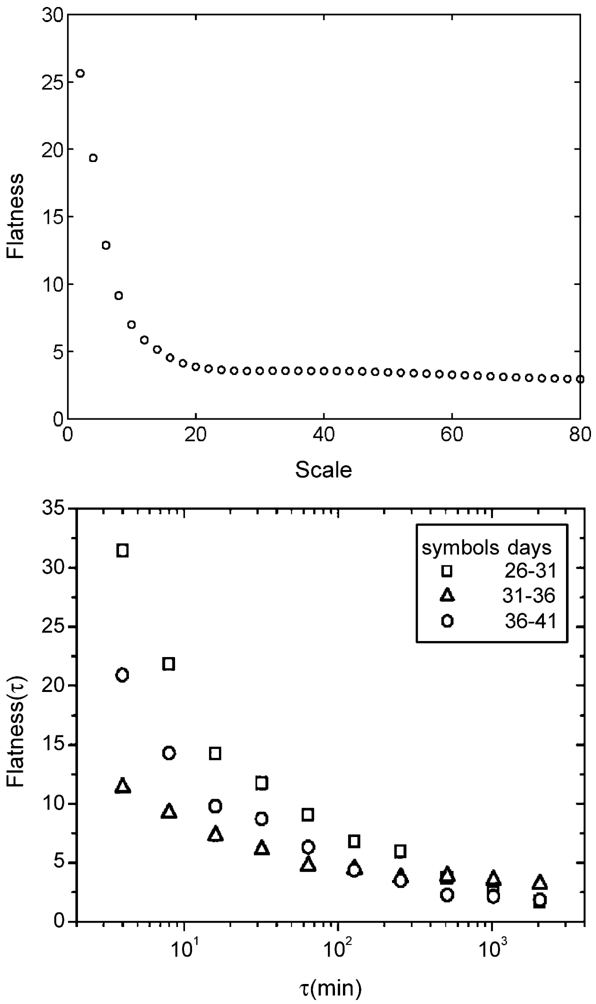
Actual fluctuations in plasma turbulence generally do not have the conventional geometrical properties. We must then devise some abstract "measure" to characterize the properties of the fluctuations and evaluate its fractal characteristics which may be interpreted with geometrical analogs. Consider, for example, the spatial series of the simulated fluctuations of the strength of the magnetic field, $B(x_i)$, along some constant value of y at time t for the two-dimensional homogeneous MHD turbulence discussed in the previous sections, where $x_i = i\delta$ with $i = 0, 1, 2, \ldots, N$ and δ is the length between grid points. From this series, we can construct a spatial series by considering, for example, the absolute value of the fluctuations due to the coarse-grained difference of the strength of the magnetic field between two spatial values $x_i + \Delta$ and x_i with $\Delta = k\delta$, some multiple of δ:

$$\delta B_i = |B(x_i + \Delta) - B(x_i)| \tag{24}$$

within the spatial interval $X = N\delta$. The simulation is statistically homogeneous over the interval X. Thus, we may calculate the ensemble average of δB_i over the interval X,

$$S_1(\delta B, \Delta) = \langle |B(x_i + \Delta) - B(x_i)| \rangle \tag{25}$$

and use it as a "measure" for the coarse-grained fluctuations of the simulated spatial series. The choice of taking



Space Plasmas, Dynamical Complexity in, Figure 12
***Top*: Flatness as a function of scale in units of grid spacing for the intermittent magnetic field fluctuations *B*(*x*) using the Haar transform for a 2D homogeneous MHD simulation at given time *t* and for fixed *y*. *Bottom*: Flatness versus time scale for three different time intervals for the solar wind during solar minimum (i. e., when the solar activity is very weak) as measured by the WIND spacecraft. Reprinted from [11] with thanks to the American Institute of Physics**

the ensemble average of the absolute values of the coarse-grained differences instead of the values of the raw differences is for the purpose of better statistical convergence [8, 127]. We can then plot $\log S_1(\delta B, \Delta)$ against $\log \Delta$ for different choices of Δ (or k). If the result approximates a straight line for some range of Δ for small Δ, we can then assign a fractal number $\zeta_1(\delta B)$ to the fluctuations of the strength of the magnetic field as:

$$\zeta_1(\delta B) = \mathrm{d}(\log S_1(\delta B, \Delta))/\mathrm{d}(\log \Delta) \tag{26}$$

for this range of small Δ. In other words, within this range of small Δ we may represent $S_1(\delta B, \Delta)$ as power of Δ with exponent $\zeta_1(\delta B)$. This exponent may be considered as an analog to the classical concept of dimension. It is generally an irrational number and usually cannot be surmised from simple dimensional analysis arguments; thus it is sort of a "fractal dimension" for the particular choice of "measure" described above.

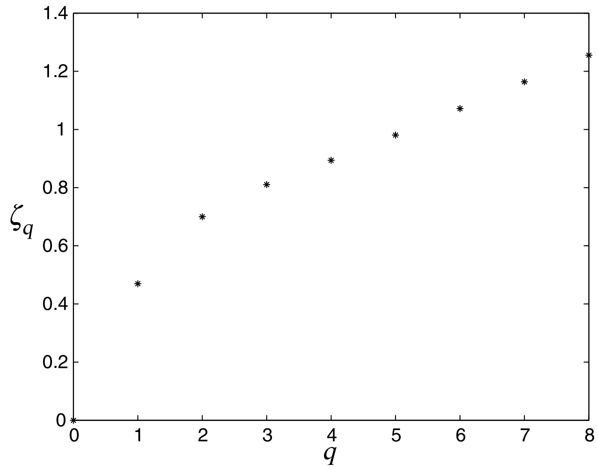
Structure Functions

It is obvious that if another choice of measure is made, the corresponding fractal number for the same spatial series will generally be different. For example, one might wish to look at higher order moments of the PDFs $P(\delta B, \Delta)$ of δB, the so-called structure functions:

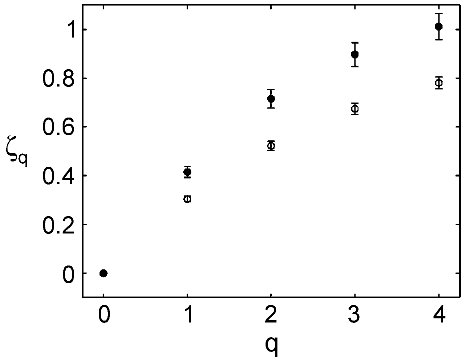
$$\begin{aligned} S_q(\delta B, \Delta) &= \int (\delta B)^q P(\delta B, \Delta)\,\mathrm{d}(\delta B) \\ &= \left\langle |B(x_i + \Delta) - B(x_i)|^q \right\rangle \end{aligned} \tag{27}$$

The motivation here is that different moments emphasize different peaks in the fluctuating series. Generally, corresponding to each ζ_q there will be a fractal exponent ζ_q for small Δ. If $\zeta_q = \zeta_1 q$, then the fractal property of the fluctuating series is fully characterized by the value of ζ_1. Such fluctuations are then said to be "monofractal" or "self-similar", i. e., the fractal characteristics for all the moment orders are similar to each other. For general intermittent turbulence, on the other hand, ζ_q would be a nonlinear function of q, or "multifractal".

In practical calculations, it has been suggested by Benzi et al. [6] for hydrodynamic turbulence studies that the correlations among the structure functions seemed to be much more apparent than those exhibited by the correlations of individual structure functions directly with coarse-grained scaling. This is particularly true for structure functions based on the absolute values of the differences of fluctuations as it is defined here [51]. This also seemed to be true for plasma turbulence studies [19,20,95, 100]. Thus, it has become very popular to first determine the relative values of the structure function exponents with respect to the structure function exponent of a particular order, and then determine the exponent for that particular order more accurately by some phenomenological, dimensional, or empirical reasoning. This procedure is called "extended similarity scaling" (ESS). Figure 13 is the result of such a statistical calculation for the fluctuations of the magnetic field strength $B(x_i)$ of the 2D homogeneous MHD simulation for given y and t. We note that the structure function exponents exhibit nonlinear, i. e., multifractal, behavior as expected for intermittent turbulence.



Space Plasmas, Dynamical Complexity in, Figure 13
Structure function exponents for the magnetic field intensity at given time *t* and for fixed *y* using ESS for a 2D MHD simulation for homogeneous intermittent turbulence. The values of the exponents at high orders are less accurate than those for the low orders due to the size limitation of the simulation box. Reprinted from [27] with thanks to the American Institute of Physics



Space Plasmas, Dynamical Complexity in, Figure 14
Structure function exponents using ESS for the time series of the strength of the magnetic field collected for extended periods during solar minimum (low solar activity) by the WIND spacecraft (*circles*) and solar maximum (high solar activity) by the ACE spacecraft (*filled circles*). Adapted from [11] with thanks to the American Institute of Physics

Calculations for structure function exponents have been carried out for magnetic and velocity fluctuations in the solar wind, e. g. [11,12,13,14,17,58,59,60,86,97,121] and elsewhere in the space environment, e. g. [57], usually by assuming the applicability of the Taylor hypothesis. We caution here again that except for most of the solar wind studies, the assumption of the Taylor hypothesis must be exercised with a critical eye. In Fig. 14, the results for such calculations for the time series of the strength of the magnetic field collected for extended periods during solar minimum (low activity) and solar maximum (high activity) by two separate spacecraft (Wind and ACE) are presented. The multifractal nature of the intermittent turbulence is apparent in both calculated results [11].

Structure function calculations may be performed conveniently for a fluctuating series for positive values of q; but will generally diverge for $q < 0$. An alternative procedure to evaluate the fractal characteristics of different moment orders is the so-called "singularity analysis" described below.

Partition Functions, Generalized Dimensions and Singularity Spectra

As we have seen in the previous discussions, turbulence in space plasmas is generally intermittent and therefore probably composed of an admixture of fluctuations of different fractal characteristics. We may attempt to extract this "multifractal" nature of the observed fluctuating time series by searching for the "dominant singular behavior" for different moment orders at small scales following the technique suggested by Parisi and Frisch [98] and Halsey et al. [54] and first applied to solar wind turbulence by Burlaga [12,13,14,15], and later for the auroral electrojet (AE) index by Consolini [41], for the high latitude geomagnetic fluctuations by Vörös [129,130], for the plasma sheet by Lui [76] and Weygand et al. [134], to the magnetospheric cusp by Yordanova et al. [139], and to the auroral electric field fluctuations by Chang et al. [34].

The basic idea here is to define an appropriate measure that is scale dependent. (Instead of referring to a spatial series such as that considered above for the numerical simulation, for variance, we consider below a generic time series based on temporal observations). To proceed, we first define an "incremental measure":

$$\delta\mu_j = \left|B(t_j+\delta) - B(t_j)\right| \tag{28}$$

where δ is the sampling interval. We now subdivide the total time interval T into $M = T/\tau$ segments with $\tau = k\delta$ and calculate the normalized scale-dependent "segmental measure"

$$\mu_i(\tau) = \sum_{j=(i-1)k+1}^{ik} \delta\mu_j/\mu \quad \text{with} \quad \mu = \sum_{j=1}^{N} \delta\mu_j \tag{29}$$

For convenience, we will generally choose M to be an integer.

The normalization of $\mu_i(\tau)$ is an attempt to represent the segmental measure as a measure of probability. This is the scale dependent measure that we seek to define. We shall assume that each such measure varies with the scale τ in a singular manner as a power law. We now form the qth moment order of the coarse-grained probabilities $\mu_i(\tau)$, traditionally called the "partition function" [54]:

$$\Gamma(q,\tau) = \sum_{i=1}^{M} \mu_i^q(\tau) \tag{30}$$

where q can be any real number. We shall now search for the dominant singular behavior of $\Gamma(q,\tau)$ as characterized by a power law in τ with exponent $\gamma(q)$ for small τ similar to that for the structure function analysis. In general, for each moment order, $\gamma(q)$ is a different number characterizing the particular fractal behavior of the subset of fluctuations, which dominates the (singular) scaling behavior of that particular moment order.

As mentioned above, Consolini et al. [41] have carried out such a singularity measure multifractal analysis for a typical time series of the auroral electrojet (AE) index, Fig. 15 (top left panel). In their calculations, the incremental measure was chosen as:

$$\delta\mu_j = \left|\phi_{\mathrm{AE}}(t_j+\delta) - \phi_{\mathrm{AE}}(t_j)\right|^2 \tag{31}$$

where $\phi_{\mathrm{AE}}(t_j)$ is the measured AE index at time t_j and δ is the sampling interval. We note from the figure that $\gamma(q)$ is indeed a nonlinear function of q, and thus, a "multifractal". There are two distinct asymptotic regimes of q ($q < -2$, and $q > +2$) for which $\gamma(q)$ varies with q nearly linearly and a nonlinear crossover region between the two asymptotes.

Henschel and Procaccia [55] suggested the concept of the so-called "generalized dimensions" for fractals. Their definition of the generalized dimension D_q is related to the partition function $\Gamma(q)$ as follows:

$$D_q = (q-1)^{-1} \lim_{\tau\to 0} \log\Gamma(q)/\log\tau = \gamma(q)/(q-1)\,. \tag{32}$$

It is called the generalized dimension because in the limit of $q \to 0$, D_0 is essentially the fractal (or similarity) dimension defined by Mandelbrot [83]. Also, it may be shown that, in the limit of $q \to 1$, D_1 corresponds to the so-called information dimension [5]. In addition, D_q's for $q \geq 2$ are the corresponding qth correlation dimensions [50]. The calculated generalized dimension D_q as a function of moment order q for the AE index is shown in Fig. 15 (top right panel). As shown by Consolini et al. [41], such a nonlinear variation may be modeled rather accurately by the so-called p-model which is a "two-scale cascade model" as follows [16,92]:

$$D_q = \log_2[p^q + (1-p)^q]^{1/(1-q)} \quad \text{with} \quad p = 0.746 \pm 0.002\,. \tag{33}$$

where p is the single fitting parameter associated with the fragmentation probability of the cascade process. This one-parameter representation of the generalized dimension for the multifractal phenomenon of the AE index reinforces the idea stated earlier that the multifractal phenomenon can be thought of as the crossover phenomenon between two invariant regions.

Another way to gauge the multifractal characteristics in terms of the partition function formalism is the method of the singularity spectrum $f(\alpha)$ [54,98]. If we associate with each segmental measure μ_i a singularity index α_i such that $\mu_i = \tau^{\alpha_i}$ where α_i is within some range between α and $\alpha + \Delta\alpha$ and pass on to the continuum limit by writing the number of differential measures whose singularity index is between α and $\alpha + \mathrm{d}\alpha$ as $\rho(\alpha)\tau^{-f(\alpha)}\mathrm{d}\alpha$, then the partition function as defined in (30) in the continuum limit is:

$$\Gamma(q,\tau) = \int \mathrm{d}\alpha' \rho(\alpha')\tau^{-f(\alpha')+q\alpha'}\,. \tag{34}$$

Since we are interested in the phenomenon for very small τ, the integral is dominated by the value of α' which makes the expression $-f(\alpha') + q\alpha'$ the smallest. Assuming $\rho(\alpha') \neq 0$ and $f(\alpha')$ smooth, the minimum is located at $\alpha' = \alpha(q)$ with

$$f'(\alpha(q)) = q \quad \text{and} \quad f''(\alpha(q)) < 0 \tag{35}$$

Therefore the partition function for very small τ is approximately

$$\Gamma(q,\tau) \sim \exp\left\{[q\alpha(q) - f(\alpha(q))]\ln\tau\right\} \tag{36}$$

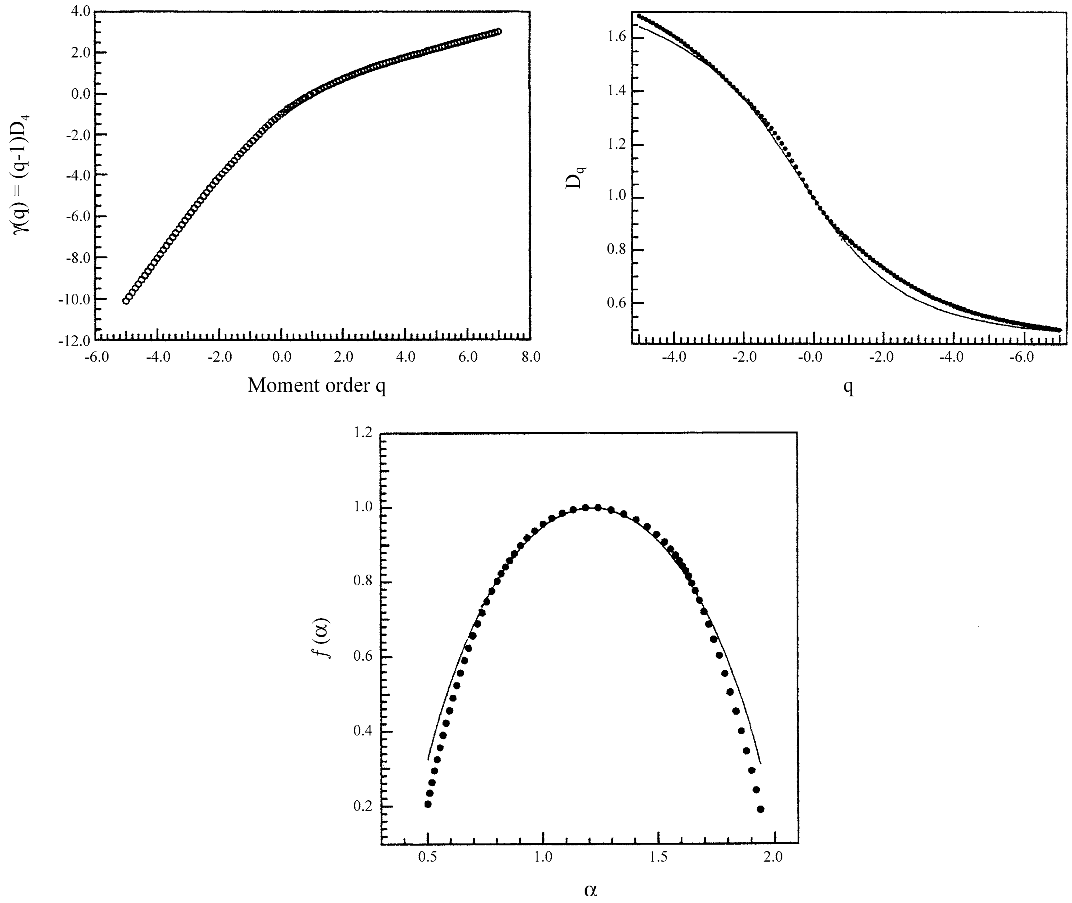
This leads immediately to the identification of

$$\gamma(q) = D_q(1-q) = [q\alpha(q) - f(\alpha(q))] \tag{37}$$

Differentiating (37) with respect to q and using the first expression of (35), we obtain

$$\alpha(q) = \gamma'(q) \equiv \mathrm{d}\gamma/\mathrm{d}q \tag{38}$$

Thus, given $\gamma(q)$, we may determine $\alpha(q)$ and then, $f(\alpha)$. Such a singularity spectrum for the AE index considered by Consolini et al. [41] is given in Fig. 15 (bottom panel). Because $\alpha(q)$ is intimately related to $\gamma(q)$ (and therefore



Space Plasmas, Dynamical Complexity in, Figure 15

***Top left*: Partition function exponent $\gamma(q)$ for a time series of the AE (auroral electrojet) index. *Top right*: Generalized dimension D_q for the same AE time series. Solid line is the best fit using the *P*-model. *Bottom*: Singularity spectrum $f(\alpha)$ for the same AE time series obtained from the Legendre transform. Adapted from [41] with thanks to the American Physical Society**

also to the partition function $\Gamma(q)$), $f(\alpha)$ sorts out the spectrum of singularities of the intermittent fluctuations, that may be mapped to the multifractal dimensions D_q as defined by (32).

In the above, we have introduced a sequence of algebraic manipulations mainly for the purpose of connecting three different popular approaches of identifying the singularity nature of the multifractal characteristics of a time (or spatial) series based on the partition function formulation. We must emphasize that these are merely different approaches of viewing fracture characteristics of intermittent fluctuations with no additional physics involved, except perhaps with the following useful thermodynamic analogy. That is, the relationship between $(\gamma(q), q)$ – or equivalently (D_q, q) – and $(f(\alpha), \alpha)$ is formally a Legendre transform. Thus, one may easily associate the relationships among such measures with the standard notion of Legendre transforms of classical thermodynamics [96]. Such a connection may become useful in identifying analogies between singularity measures of multifractals with multicriticality measures in thermodynamics.

It is interesting to note that in multifractal analyzes of space plasma turbulence in regions where intermittency are expected, similar functional dependence of $\gamma(q)$ on q as above is also generally obtained [12,13,14,15,18,34,87,134,139]. Although, we must recall that for turbulence in the plasma sheet or the auroral zone and even sometimes in the solar wind, the assumption of the validity of the Taylor hypothesis must be applied with caution.

In the previous section, it was demonstrated that intermittency characteristics of a fluctuating series could be brought out using the technique of wavelet transforms. Certain quantities, Local Intermittency Measure LIM or Normalized Power and Flatness, were introduced using the wavelet transforms to gauge the intermittency effects. Since the complete set of coefficients of the wavelet transforms essentially represents the full fluctuating event, it is reasonable to inquire if it can be used to advantage to describe the full characteristics of intermittency of a fluctuating series. The answer is a definite *yes*. For example, wavelet transforms have been used to study multifractal characteristics of time series related to intermittent turbulence in space plasmas by Yordonova et al. [139] for the magnetospheric cusp and Vörös et al. [131] for the plasma sheet.

Rank-Ordered Multifractal Analysis (ROMA)

In the above, intermittent fluctuations were analyzed using the structure function and/or partition function methods based on the statistics of the full set of fluctuations. Since most of the observed or simulated intermittent fluctuations are dominated by fluctuations with small amplitudes, the subdominant fractal characteristics of the minority fluctuations – generally of larger amplitudes – are easily masked by those characterized by the dominant population. It therefore appears useful to search for a procedure that explores the fractal nature of the subdominant fluctuations by first appropriately isolating out the minority populations and then perform statistical investigations for each of the isolated populations.

Phrasing it differently, for intermittent fluctuations exhibiting multifractal characteristics, we visualize that these fluctuations are composed of many types, each type being characterized by a particular fractal dimension. The questions to ask then are (i) what are the different fractal dimensions? and (ii) how are the various types or classes of fluctuations distributed within the turbulent medium? We answer these questions quantitatively by constructing a "rank-ordered multifractal spectrum".

Using the example of a 2D MHD simulation that was discussed in Sect. "Non-Gaussian Probability Distribution Functions", we demonstrate below how this idea may be achieved with a rank-order method that subdivides the fluctuations into groupings based on the range of the scaled-sizes of the fluctuations [28]. In the simulation, ideal compressible MHD equations expressed in conservative forms are solved numerically with 1024 × 1024 grid points in a doubly periodic (x, z) domain of length 2π in both directions using the WENO code [64] so that the total mass, energy, magnetic fluxes and momenta are conserved. The initial condition consists of random magnetic field and velocity with a constant total pressure for a high beta plasma. As discussed previously, after sufficient elapsed time, the system evolves into a set of randomly interacting multiscale coherent structures exhibiting classical aspects of intermittent fluctuations.

We recall that the PDFs $P(\delta B^2, \varDelta)$ of the square of the magnitude of the magnetic field δB^2 for such intermittent fluctuations were non-Gaussian and became more and more heavy-tailed at smaller and smaller scales (Fig. 8). An attempt to collapse the unscaled PDFs according to the one-parameter scaling formula (12) re-displayed below as (39):

$$P(\delta B^2, \delta)\delta^s = P_s(\delta B^2/\delta^s) \tag{39}$$

indicated approximate scaling with an estimated scaling exponent of $s = 0.335$ (Fig. 9). The scaling formula (39) may be interpreted from the concept of fractals. We begin the interpretation by assuming that δB^2 and P vary with δ as power laws: $\delta B^2/\delta^a = I$ and $P/\delta^b = J$, where (a, b) are the power-law (fractal) exponents and (I, J) are constants – i. e. invariants with respect to the scale δ. If the form of $P(\delta B^2, \delta)$ is invariant as the scale changes, then it has been shown that a functional relation exists between the two invariants, (I, J) [29]. Imposing the normalization condition for the PDFs, we obtain the one-parameter scaling form (39) with $s = a = -b$.

Such a phenomenon is self-similar and the PDFs collapse onto one scaled PDF $P_s(Y)$ where $Y = \delta B^2/\delta^s$ is a scale invariant. In such a case, the structure function exponents for the PDFs will satisfy the monofractal property of $\zeta_q = sq$. The scaling exponent s may be interpreted as a single fractal measure that characterizes the fluctuations of all scales through the scaling relation (39). If the PDFs are self-similar Gaussian distributions for all scales, representing random diffusion, the scaling exponent s is equal to 0.5. For other monofractal distributions, the scaling exponent may take on any positive real value.

For the current example, structure function calculations based on the full set of simulated fluctuations showed a nonlinear relation between the exponents and the moment order, indicating the phenomenon is multifractal in nature [27]. However, the physical interpretation of the multifractal nature is not easily deciphered by merely examining the curvature of the deviation from linearity.

We shall therefore attempt to perform statistical analyzes individually for subsets of the fluctuations that characterize the various fractal behaviors within the full multifractal set. Such grouping of fluctuations must depend somehow on the sizes of the fluctuations. However, the groupings cannot depend merely on the raw values of the sizes of the fluctuations because the ranges will be different for different scales. Thus, we proceed to rank-order the sizes of the fluctuations based on the ranges of the scaled variable $Y = \left|\delta B^2\right|/\varDelta^s$ where s is the scaling exponent for each grouping defined in (39). (Absolute values of the fluctuations are used to take advantage of the symmetry property of the PDFs and for the purpose of better statistical convergence.) For each chosen range we shall assume that the fluctuations of all scales will exhibit monofractal behavior satisfying the form of (39) and be characterized by an scaling exponent s. The question is then how can this procedure be accomplished.

We begin by considering a differential range of scales dY in the vicinity of some scaled size $Y = \left|\delta B^2\right| / \Delta^s$. Fluctuations whose sizes fall within such a differential range will probably exhibit monofractal behavior characterized by some scaling exponent s such that the differential structure function $\mathrm{d}S_q$ will vary with the scale as Δ^{sq}, i. e.,

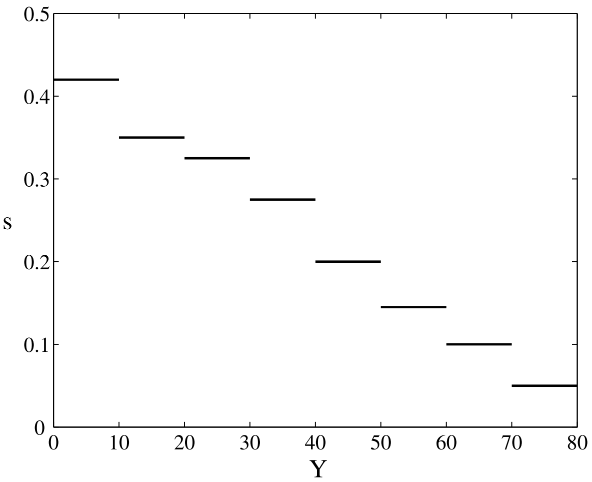
$$\mathrm{d}S_q \triangleq (|\delta B^2|)^q P(|\delta B^2|, \Delta)\,\mathrm{d}(|\delta B^2|) = \Delta^{sq} Y^q P_s(Y)\mathrm{d}Y \tag{40}$$

Given a set of PDFs $P(|\delta B^2|, \Delta)$, the corresponding multifractal spectrum $s(Y)$ may be obtained approximately (if the ansatz is valid) by integrating the functional differential expression (40) over small contiguous ranges of ΔY with the assumption that within each incremental range the scaling exponent s is essentially a constant [28]. Thus, for a generic range of ΔY within (Y_1, Y_2), we form such a range-limited structure function as follows:

$$\Delta S_q(\left|\delta B^2\right|, \Delta) = \int_{a_1}^{a_2} \left|\delta B^2\right|^q P(\left|\delta B^2\right|, \Delta)\,\mathrm{d}\left|\delta B^2\right|$$

$$\simeq \Delta^{sq} \int_{Y_1}^{Y_2} Y^q P_s(Y)\mathrm{d}Y \tag{41}$$

where $a_1 = Y_1\Delta^s$ and $a_2 = Y_2\Delta^s$. We then search the value of s such that the scaling property of the range-limited structure function that varies with s is $\Delta S_q(s) \sim \Delta^{sq}$. If such a value of s exists, then we have found one region of the multifractal spectrum of the fluctuations such that the PDFs in the range of ΔY collapses onto one scaled PDF. Performing this procedure for all contiguous ranges of ΔY produces the rank-ordered multifractal spectrum $s(Y)$ that we are looking for. The determined value of s for each grouping should be un-affected by the statistics of other subsets of fluctuations that are not within the chosen range ΔY and therefore should be quantitatively quite accurate. The physical meaning of this spectrum is that the PDFs for all time lags collapse onto one master multifractal scaled PDF. The spectrum is implicit since Y is a defined as a function of s.



Space Plasmas, Dynamical Complexity in, Figure 16
Rank-ordered multifractal spectrum, $s(Y)$. $Y = |\delta B^2|/\Delta^s; |\delta B^2|$ in units of bin size and Δ in grid spacing. The spectrum is calculated for 8 contiguous ranges of ΔY. Reproduced from [28] with thanks to the American Physical Society

Interestingly, for this particular example, there exists one and only one value of s in each range of ΔY, that satisfies the above constraint. Unlike the structure functions defined for the full range of fluctuations, the range-limited structure functions based on (41) exists also for negative real values of q (except for the range including $Y = 0$). Figure 16 displays the calculated rank-ordered spectrum $s(Y)$ based on eight contiguous ranges of ΔY [28]. It is noted that the spectrum has values of s ranging between 0.5 and 0.0. The spectrum can be refined by choosing more range intervals with smaller range sizes of ΔY, although in practice this procedure is limited by the availability of simulated data points. At $Y = 0$, the scaling exponent s appears to approach the self-similar Gaussian value of 0.5 representing random diffusion. As the value of the scaled fluctuation size Y increases the scaling exponent decreases accordingly indicating the fluctuations are becoming less and less space filling. At very large value of Y, the scaling exponent s seems to asymptotically approach the value of zero indicating the fluctuations have become extremely sparse.

Such an implicit multifractal spectrum has several advantages over the results obtainable using the conventional structure function calculations. Firstly, the utility of the spectrum is to fully collapse the unscaled PDFs. Secondly, the physical interpretation is clear. It indicates how space-filling (in terms of the value of s) are the scaled fluctuations once the value of Y is given. Thirdly, the determination of the values of the fractal nature of the grouped fluctuations is not affected by the statistics of other fluctuations that do not exhibit the same fractal characteristics. Fourthly, it provides a natural connection between the one-parameter scaling idea (39) and the multifractal behavior of intermittency.

Invariant Scaling

Forced and/or Self-Organized Criticality (FSOC)

In the above sections, we provided some convincing arguments as well as numerical and observational evidences indicating that space plasma turbulence is generally in a state of dynamic "topological complexity". By "complex" topological states we mean magnetic topologies that are not immediately deducible from the basic (e. g., MHD) equations. We also discovered that the structure/partition functions for turbulence generally scale as power laws at least at small scales. Expressed another way, we found that there were power-law scale invariants with respect to coarse-graining at small scales where intermittency prevails. Below, we shall briefly address the salient features of such "topological phase transitions" as well as the associated scale invariant properties in turbulent plasmas. (We note here that the concept of phase transitions has also been employed to address the phenomenon of complexity in space plasmas by Sitnov et al., [106] using the singular spectrum method.)

For nonlinear stochastic systems exhibiting complexity, the correlations among the fluctuations of the random dynamical fields (electric, magnetic and velocity fields, etc.) are generally very long-ranged and there exist many correlation scales (as exhibited by the various outsized plasma coherent structures). The dynamics of such systems are notoriously difficult to handle either analytically or numerically. On the other hand, since the correlations are extremely long-ranged, it is reasonable to expect that the system will exhibit some sort of invariance property under coarse-graining scale transformations. A powerful analytical technique that exploits this invariance property is the method of the dynamic renormalization group (DRG) which we shall briefly describe below. The technique is a generalization of the static renormalization group introduced by Wilson in 1972 [135] for equilibrium phase transitions. Dynamic renormalization group was originally developed as a perturbation theory by Halperin et al. [53] and first applied to hydrodynamic problems by Forster et al. [49]. A closed form exact theory of DRG was developed by Chang et al. [30]. And a review of the basic elements of this theory may be found in Sect. "Non-Gaussian Probability Distribution Functions" of a Physics Report by Chang et al. [31].

No matter how complex the stochastic behavior is for the turbulent system, we may conjecture that the system is somehow characterized by a set of dynamically evolving parameters $\{P_n\}$. We now transform the stochastic system by coarse-graining and ask how these parameters transform accordingly (subject to the rules stipulated by the underlying dynamical equations, e. g., those for the interacting flux tubes). Symbolically, we write

$$\partial\{P_n\}\partial\ell = R\{P_n\} \tag{42}$$

where R represents the coarse-graining (renormalization-group) transformation operator and ℓ *is* the coarse-graining parameter. The operator is generally very complicated for even relatively simple stochastic systems. Instead of actually delving into the mathematical formalism of DRG, we shall merely explore some of the basic concepts related to such ideas.

Forced and/or Self-organized Criticality Generally, if one starts at some initial state $\{P_n(0)\}$ and proceeds with the coarse-graining procedure following the prescription dictated by (42), the result would be a trajectory in the phase space characterized by the parameters. In analogy to nonlinear dynamics, we expect that there may be fixed points (singular points) in the phase space flow field, at which $d\{P_n\}/d\ell = 0$. At such a point, e. g., $\{P_n^*\}$, the correlation length should not be changing. However, the coarse-graining transformation requires that all length scales to change with ℓ. Therefore, to satisfy both requirements, the correlation length must be either infinite or zero at the fixed point. When it is at infinity, the state of the dynamical system would then be at a state of dynamic criticality, analogous to the state of criticality in equilibrium phase transitions.

Generally, a dynamical system will not be near a parametric state such as $\{P_n^*\}$. But, if the coarse-graining procedure (i. e., the renormalization group trajectory) leads it eventually to a point in the parametric space close to a fixed point such as $\{P_n^*\}$, then in a coarse-grained sense the stochastic system would be close to a state of dynamic criticality. We may then approximate the coarse-graining transformation (42) by linearizing the operator R. The result is a set of coupled linear differential equations characterizing the variations of the deviations of the parameters $\{P_n\}$ from $\{P_n^*\}$.

$$\mathrm{d}\{Q_n\}/\mathrm{d}\ell = R_\mathrm{L}\{Q_n\} \tag{43}$$

where $Q_n = P_n - P_n^*$ and R_L is the linearized renormalization-group matrix operator.

We assume that expression (43) may be expressed in terms of a set of eigenvalue equations:

$$\mathrm{d}\mu_i/\mathrm{d}l = \lambda_i \mu_i \qquad (44)$$

where the eigenvectors μ_i's are linear combinations of the parameters Q_n and the λ_i's are the corresponding eigenvalues. We shall assume that the eigenvectors are complete and the eigenvalues are all real. (Nondiagonalizable operator R_{L} may render the eigenvectors incomplete and lead to logarithmic corrections to the discussion below, which will not be considered here).

The solutions of (44) are simply $\mu_i = \mu_i^0 \exp \lambda_i \ell$ where $\{\mu_i^0\}$ represents some arbitrary state that is close to $\{P_n^*\}$. From this, it can be verified immediately that

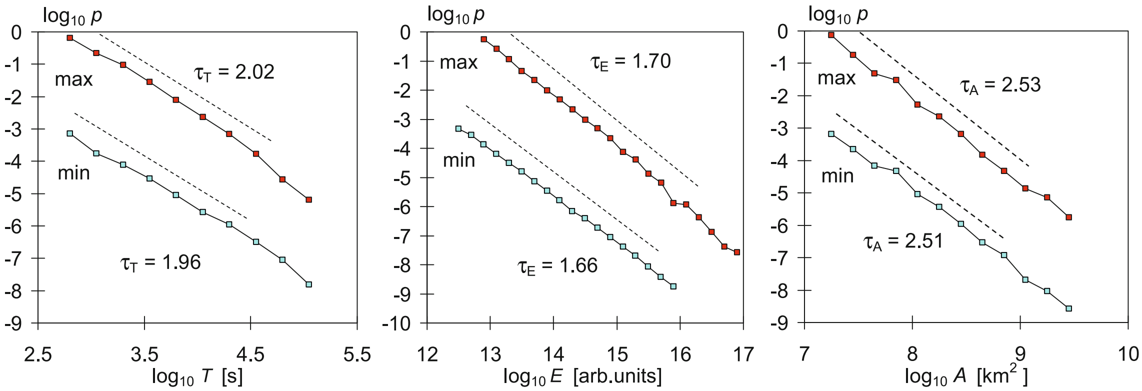
$$\mu_i/\mu_j^{\lambda_i/\lambda_j} = \text{constants} \qquad (45)$$

i. e., the μ_i's form power law invariants which are independent of the coarse-graining scale transformation ℓ.

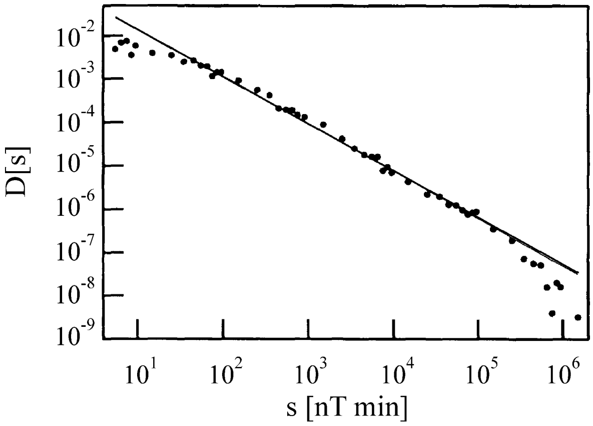
Expressing this result in words: The mathematical consequence of the linear approximation is that, close to dynamic criticality, certain linear combinations of the perturbations of the parameters that characterize the stochastic phenomenon will correlate with each other in the form of power laws. These include the (k, ω), i. e. mode number and frequency, spectra of the correlations of the various fluctuations of the dynamic field variables in space and time, etc. The fact that the dynamical (stochastic) system must initiate from some state so that the coarse-graining trajectory will carry it to the close proximity of a fixed point such as $\{P_n*\}$ means that there may need some initial tuning of the state for the system to have such behavior. For this reason, we shall refer to such dynamical systems as systems near Forced and/or Self-organized Criticality (FSOC) [22]. The phenomenon of FSOC is spatiotemporal and the invariant scaling idea permeates throughout the time-space domain. In addition, the scaling between time and space generally is not isotropic.

Such power law behavior has been detected in probability distributions related to the solar flares [74,126], Fig. 17, in the AE burst occurrences as a function of the AE burst strength [37], Fig. 18, in the global auroral UVI imagery of the statistics of size and energy dissipated by the magnetospheric system [81], in the probability distributions of spatiotemporal magnetospheric disturbances as seen in the UVI images of the nighttime ionosphere [124], in the probability distributions of durations of Bursty Bulk Flows [2], and in the magnetic field fluctuations of the Earth's magnetotail [61,93]. Similar behavior has also been obtained in numerical simulations for the distributions for solar flares [45], for dynamical models for the current sheet [66,110,122] and other applications. This invariant behavior had been called self-organized criticality (SOC) [3], i. e., tuning is generally not necessary.

An eigenvector (eigen-parameter) μ_i is "irrelevant" if the corresponding eigenvalue λ_i is less than zero because in the limit of large ℓ the value of μ_i will become negligible. The reverse is true for positive λ_i's. We shall not consider the special case of $\lambda_i = 0$ which introduces singularities of fractional powers of logarithms. If there are only a small number of relevant eigenvalues with $\lambda_i > 0$, the corresponding FSOC system has only a small number of "relevant parameters" that are needed to characterize the stochastic state, i. e., the system may appear to have low dimensionality in parameter space near criticality [22].



Space Plasmas, Dynamical Complexity in, Figure 17
Flare probability distributions of avalanche lifetime (*left*), emission flux (*middle*) and peak area (*right*) for solar max (high activity) and min (low activity) obtained for emissions detected by the EIT (extreme ultraviolet imaging telescope) aboard the SOHO spacecraft in the 195 Å wavelength above a threshold of 40% of the average emission. The min distributions are shifted downward for comparison. Adapted from [126] with thanks to the American Physical Society



Space Plasmas, Dynamical Complexity in, Figure 18
Distribution $D(s)$ of the burst size $s = \int_{\Omega}[AE(t) - L]dt$ of an AE index time series, where the integration is performed over the time interval Ω when the quiet level of $L = (45 \pm 15)nT$ is exceeded. The solid line is a power-law best fit. Reprinted from [37] with permission from the Italian Society of Physics

From Criticality to Intermittency

We now ask a deeper question: How does the invariant behavior of the structure or partition functions, which changes with moment order for intermittent turbulence amalgamate (reconcile) with the concept of criticality (SOC or FSOC) outlined above?

For systems near criticality the correlation functions are generally related to the correlation time and space as power laws. Since the structure functions are related to the spatial correlation functions [7], it is therefore not surprising that for systems near SOC/FSOC, the structure function exponents for intermittent turbulence are power law scale-invariants in coarse-graining scale transformations.

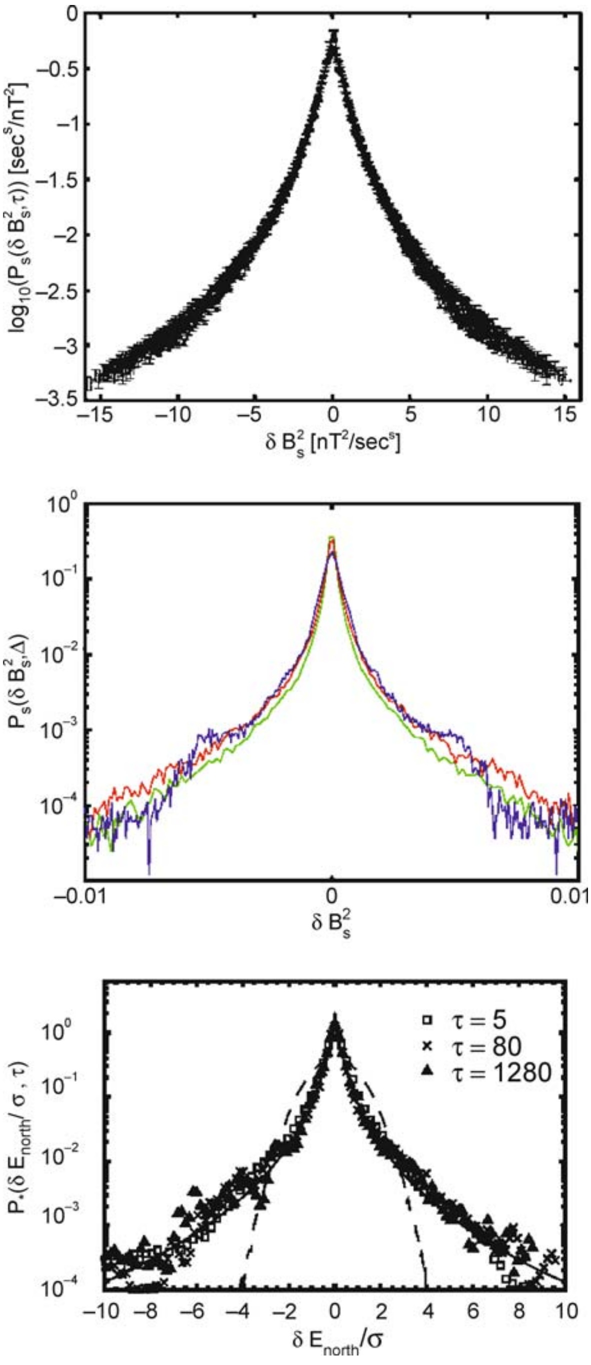
For a stochastic process that is monofractal, the structure function exponents vary with the moment order linearly. Thus a single invariant, such as $I_1 = S_1/\Delta^{\zeta_1}$ would suffice to characterize the behavior of fully developed turbulence that is self-similar. For example, it was demonstrated in Subsect. "Rank-Ordered Multifractal Analysis (ROMA)", that if the probability distributions functions (PDFs) satisfy the one-parameter scaling law of (39), the structure function exponents $\zeta_q = sq$ where s is the scaling exponent, and the PDFs collapse onto one master scaling function $P_s(\delta B^2/\Delta^s)$. This is approximately satisfied by the PDFs for the numerical simulation results discussed in Sect. "Non-Gaussian Probability Distribution Functions" [33] (Fig. 9 and re-displayed here as Fig. 19, middle panel). Similar results have been found by Hnat et al. [56] for their study of the PDFs of solar wind turbulence as seen by the WIND spacecraft and Tam et al. [112] for the electric field intensity fluctuations in the auroral zone as detected by the SIERRA rocket experiment (Fig. 19, top and bottom panels). However, it is to be noted that in all three scaling plots the deviations from the mean become larger for larger fluctuations, which are indications that multifractal characteristics will still need to be addressed. This was demonstrated in terms of the rank-ordered multifractal spectrum as discussed in Subsect. "Rank-Ordered Multifractal Analysis (ROMA)".

We note that generally – e. g., Fig. 8 – the portions of the PDFs for small fluctuations for most of the observed and simulated intermittent turbulence have shapes very close to that of a Gaussian. This is the reason why the calculated exponent s for small scaled fluctuations is close to the value of 0.5 (characteristic of self-similar random diffusion). In the numerical example considered in Subsec. "Plasma Resonances and Coherent Structures", $s \approx 0.42$ for the range of ΔY between $(0, 10)$. And the behavior of the fluctuations within this range was shown to be essentially monofractal. This would have given the value of the first structure function exponent $\zeta_1 = 0.42$ and the related power spectrum index would therefore be approximately $1 + \zeta_2 = 1 + 2\zeta_1 \approx 1.84$ (Frisch, 1995). If no attempt were made in rank-ordering and the value of s were determined for the full set of fluctuations by looking for power-law behavior of $P(0, \Delta) \sim \Delta^s$, the best value was found to be approximately 0.335. Assuming approximate monofractal scaling, this number would give a power spectrum index of approximately 1.67. Both numbers are not very different from that obtainable using classical arguments for hydrodynamic or MHD turbulence.

Most observational or numerical "critical exponents" of SOC are obtained either using first order box counting or from calculations of density, optical, and other field intensities such as electric, magnetic, velocity, and energy fields, which generally involve statistical averages over the zeroth, first, or second (i. e. low) order moments of their related PDFs and therefore each represents an approximate monofractal behavior of some sort of weighted average of the fractal values for small scaled fluctuations.

In fact, if the experimental or numerical evaluations are based on rough estimates, the values of the exponents will generally take on essentially the Gaussian – i. e., mean-field – values. And such observed "critical exponents" would satisfy the standard scaling relations. We suggest that this is the reason why recent observations [126] and numerical simulations [69] seem to indicate the coexistence of the phenomenon of SOC and multifractal signatures for intermittent turbulence. See, e. g. Figs. 17 and 20.

Thus, based on the example of rank-order procedure of the multifractal behavior of intermittent fluctuations



◄ Space Plasmas, Dynamical Complexity in, Figure 19

***Top*: Scaled PDF based on the one-parameter scaling form of (39) with $\zeta_1 = s \approx 0.42$ for the fluctuations of the magnetic field energy density as seen by the WIND spacecraft. The data shown correspond to τ between 46 seconds and 26 hours. Adapted from [56] with thanks to the American Geophysical Union. *Middle*: Scaled PDF for magnetic field energy density fluctuations based on the one-parameter scaling form of (39) with $\zeta_1 = s \approx 0.335$ from a 2D MHD simulation for homogeneous turbulence for $\Delta = 2$ (*green*), 8 (*red*), 32 (*blue*) units of grid spacing. Adapted from [33] with thanks to the American Institute of Physics. *Bottom panel*: Normalized PDF for fluctuations of transverse electric field intensity E_{north}/σ at $\tau = 5, 80, 1280$ ms, where $\sigma = \tau^{\zeta_1}(\zeta_1 \approx 0.8)$ is the root mean square of the fluctuations. The solid line corresponds to the normalized Castaing distribution with $\lambda = 1.02$. The results of the analyzes for Gaussian fluctuations (*dashed*) are shown for comparison. Reprinted from [112] with thanks to the American Geophysical Union**

described in Subsect. "Rank-Ordered Multifractal Analysis (ROMA)", we suggest that power-law relations of SOC can be made more precise if the exponents are determined more meticulously using rank-ordered fluctuations. For each rank-order, particularly for the rank-order of the smallest scaled fluctuations, one can determine a corresponding set of more accurately determined "critical exponents" that satisfy the classical exponent relations.

In summary, the conventional concepts of SOC or FSOC as tested by the lower order spatiotemporal statistical averages would generally be satisfied for intermittent turbulence. However, the ideas of invariant scaling and scaling relations will need to be generalized to include the "crossover effects" for large scaled fluctuations.

Dynamical Modeling – The Lu–Klimas Magnetic Field Reversal Model

By utilizing results obtained from direct numerical simulations, investigators were able to confirm a number of the theoretical predictions of intermittent turbulence and complexity phenomena in space plasmas that could arise from the process of merging and interactions of plasma coherent structures. The system sizes of direct numerical simulations, however, are generally limited by the capabilities of the present day computers. To circumvent this limitation, an alternative approach to study the intermittency in space plasma turbulence numerically would be to mimic the coarse-grained dissipation and the development of complexity by means of dynamical modeling which, in turn, may lead to a better theoretical understanding of the fundamental underlying stochastic processes.

While there exists a number of dynamical models that mimic avalanche processes exhibiting scaling and some also intermittency, only a few have attempted to incorporate the basic physics of plasmas in their formulations. From these, we select one simple model, the Lu–Klimas fast-field annihilation model of magnetic field reversal, for further discussion and comments. The model has the basic ingredient of a diffusive equation [66]:

$$\frac{\partial B_x}{\partial t} = \frac{\partial}{\partial z}\left[D(z,t)\frac{\partial}{\partial z}\right]B_x + S(z,t)\,. \tag{46}$$

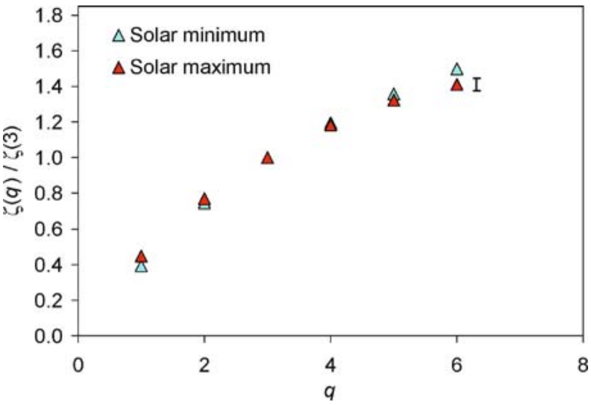
It may be obtained through the drastic reduction of the resistive MHD system in an idealized one-dimensional limit. In (46), $B_x(z,t)$ is the magnetic flux and $S(z,t)$ is a source term which could be interpreted as that characterizes the convection of the magnetic flux, although in discussions below it will be arbitrarily prescribed and assumed to be independent of time. The diffusion coefficient $D(z,t)$ is linearly proportional to an anomalous resistivity with a hysteresis characteristic as stipulated below:

$$\frac{\partial D}{\partial t} + \frac{D}{\tau} = \frac{Q(|\partial B_x/\partial z|)}{\tau} \tag{47}$$

where $Q = D_{\min}$ (low state) for $|\partial B_x/\partial z| < k$, and $Q = D_{\max} \gg D_{\min}$ (high state) for $|\partial B_x/\partial z| > \beta k$ with $\beta < 1$. Thus, Q is double-valued and dependent upon the history for $\beta k < |\partial B_x/\partial z| < k$. The value k may be thought of as a stability threshold which Klimas et al. [66] attributed to some sort of current driven instability (e. g., Lui et al. [79]). The value of Q remains in the low state ($= D_{\min}$) until the threshold is reached and switches to the high state ($= D_{\max}$). It stays at the high state until $|\partial B_x/\partial z| < \beta k$. This anomalous diffusive model is essentially that originally suggested by Lu [73], which was motivated by his studies related to the phenomena of self-organized criticality [3]; in particular, those related to the probability distributions of solar flares [74]. Unlike the Klimas' model, the source term in Lu's original formulation was assumed to be random and the boundary conditions were chosen to allow transport through the boundaries.

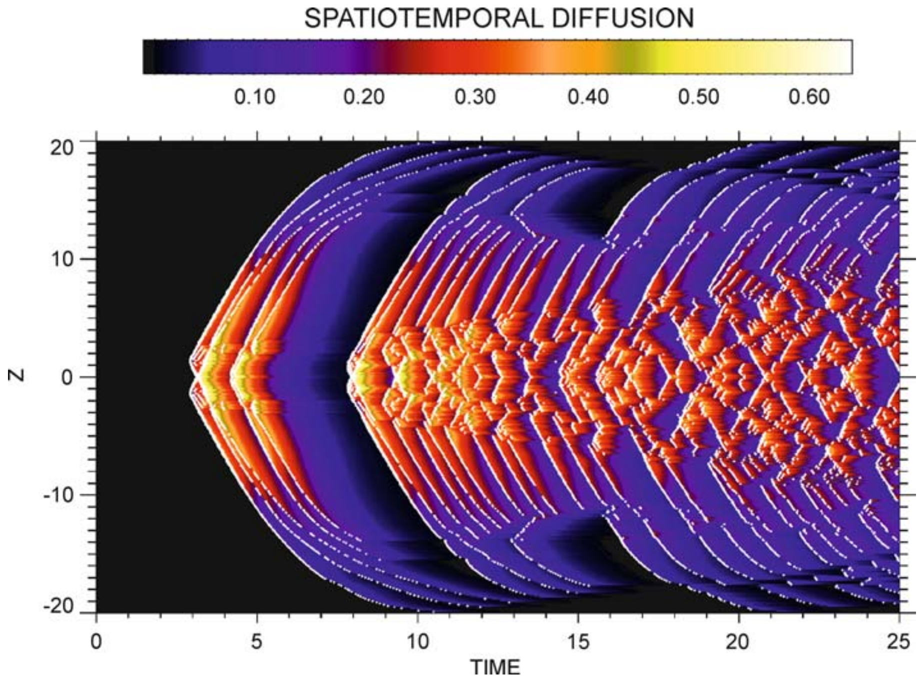
Klimas et al. [66] and Uritsky et al. [123] have studied their model numerically on a spatial interval $-L \le z \le L$ subject to the boundary conditions $\partial B_x/\partial z = 0$ at $z = \pm L$, i. e., no transport of B_x through the boundaries. They considered antisymmetric solutions with a field reversal at $z = 0$ and a time-independent source term $S(z,t) = S_0 \sin(\pi z/2L)$ which may be considered as a steady convection of B_x into the field reversal region. Setting $\tau = 1$ as the fundamental time scale and $L = 20$, they found that there exists a range of choices of parameters for which the dynamical model evolves in a sequence of quasi-periodic loading ($\mathrm{d}E/\mathrm{d}t > 0$)-unloading ($\mathrm{d}E/\mathrm{d}t < 0$) cycles where the total field energy $E(t)$ is defined by:

$$E(t) \equiv \int_{-L}^{L} \mathrm{d}z B_x^2(z,t)\,. \tag{48}$$

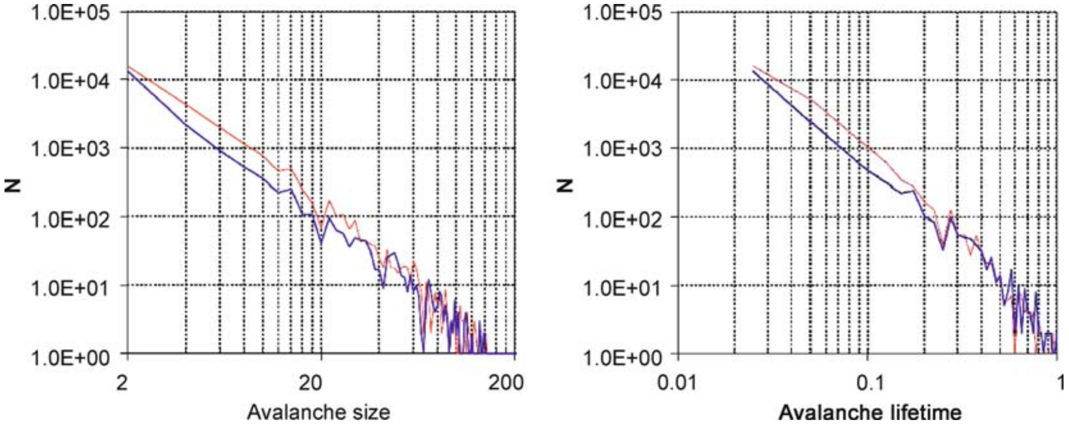


Space Plasmas, Dynamical Complexity in, Figure 20
Relative structure function exponents with respect to the third order exponent using ESS for flare statistics at solar maximum (high activity) and solar minimum (low activity) based on emissions detected by the EIT (extreme ultraviolet imaging telescope) aboard the SOHO spacecraft in the 195 Å wavelength. Adapted from [126] with thanks to the American Physical Society

During the loading intervals, when the system is quiet, the slow increase of $E(t)$ is the sole evolutionary feature of the system. As loading continues, the field gradient $|\mathrm{d}B_x/\mathrm{d}z|$, i. e. the current density, reaches the critical level for Q somewhere in the field distribution (usually at a pair of points symmetrically placed near $z = 0$). Subsequently a complicated process of unloading ensues, which is best understood in terms of the spatiotemporal evolution of the diffusion coefficient $D(z,t)$ as depicted in Fig. 21 during the initial stage of development of the unloading process, where the color scale of the figure indicates the intensity of the diffusion coefficient. In the figure, a white dot is placed at each point where the critical level of $|\mathrm{d}B_x/\mathrm{d}z|$ – or current density – for Q is reached and $Q(z,t) = D_{\max}$. As the instabilities spread and propagate toward both the center and the edges of the field distribution, more current sheets (i. e., sheet regions with strong current density) are excited and their topology becomes increasingly complex. The current sheets generate diffusion and the diffusion coefficient $D(z,t)$ is at a local maximum just after the passage of a current sheet at any given position. This strong localized diffusion enhances the transport of the magnetic flux toward $z = 0$ where fluxes of opposite signs meet and annihilate. This process of generation of complicated current sheet topology, transport of magnetic flux and annihilation adjusts self-consistently with the imposed source function such that there is a dynamically pseudo-stationary state. And, the loading-unloading process repeat quasi-periodically after an initial transient.



Space Plasmas, Dynamical Complexity in, Figure 21
Spatiotemporal evolution of the diffusion coefficient $D(z, t)$ for the Lu–Klimas model during the initial stage of development of an unloading process. The *color scale* of the figure indicates the intensity of the diffusion coefficient. A *white dot* is placed at each point where the critical level for Q is reached and $Q(z, t) = D_{\max}$. Reprinted from [66] with thanks to the American Geophysical Union



Space Plasmas, Dynamical Complexity in, Figure 22
Avalanche size (*left*) and lifetime (*right*) distributions for the Lu–Klimas model for early (*blue*) and late (*red*) intervals during a single unloading event. Reprinted from [123] with permission from Elsevier

The above phenomenon of continued sporadic generation of unstable current sheets and the accompanying broadening of the diffusive layer during the unloading process is reminiscent of the resonance broadening phenomenon of the "apparent reconnection" process in the 2D MHD simulation result for the sheared magnetic field geometry discussed earlier. Klimas et al. [67,68,69], Uritsky et al. [125] and Uritsky and Klimas [122] have recently extended their dynamical modeling effort based on the anomalous resistivity idea with hysteresis to the 2D geometry and the similarity between their calculated results of sporadic localized reconnections and the direct 2D MHD simulation results of resonance-broadening and coarse-grained dissipation – as well as the ensuing magnetic reconfiguration process – due to the sporadic, localized merging of coherent structures is even more noticeable, even though the boundary and initial conditions as well as the presumed instabilities of both sets of calculations were quite different. The main point to note is that the phenomenon of coarse-grained dissipation in the results of direct numerical simulation is replaced in the dynamical model by the assumed anomalous resistivity with hysteresis. Thus, other than the arbitrariness of the dissipation mechanism, such type of dynamical modeling has almost achieved its purported aim; i. e., using a much simplified dissipation model, large scale simulations of complexity of space plasmas may perhaps be achieved with relatively small scale computers.

However, the existing dynamical models may be too crude to lead to quantitatively comparable results of actual observations. For example, in the Lu–Klimas model, the critical threshold was arbitrary and kept at a constant value. In addition the phenomenon of hysteresis was again quite arbitrary and the double-valuedness of Q was also triggered at arbitrary levels. In the coarse-grained dissipation scenario acquired from, for example, direct numerical simulations, there was actually no specific threshold for triggering the "fluctuation induced nonlinear instability" and the dissipation during "apparent reconnection" was due to the combination of wave (e. g., Alfvén wave) propagation and resonance broadening. The hope is that perhaps even with some of such arbitrary ad hoc assumptions in the present and/or future modeling of local dissipation, the hypothesized model(s) will still lead to some of the global features such as the scale-invariant properties that will fall in the same universality class as the actual physical phenomena.

We now proceed to discuss the invariant properties of the simulated result of the one-dimensional magnetic field reversal and annihilation model. There are at least two sets of statistics that may shed light on the scale free invariant properties of SOC for the loading-unloading events that emerge from the numerical calculations of the Lu–Klimas model. Firstly, after an initial transient, the sequence of burst-like unloading events seem to occur by clustering around an average field strength, i. e., on average, the increasing strength of the field reversal is more-or-less balanced by the annihilation of the magnetic field in the current sheet. Thus, we expect that the unloading events exhibit power law distributions in, for example, energy dissipation and burst duration. These results were essentially uncovered by Klimas et al. [66] for the special situation of very low driving rates and $D_{\min} = 0$.

We may also ask the question about the scale free invariant behavior for the complicated current sheet distributions triggered by the instabilities associated with the anomalous resistivity for a single unloading event. Such small scale fluctuations again seem to self-consistently balance between the anomalous transport and field annihilation. Thus, we may again expect certain universal SOC behavior for these small scale fluctuations. To search for such behavior, Uritsky et al. [123] defined avalanches in terms of the size and duration of contiguous unstable grid points in the spatiotemporal domain. Power law statistics of the probability distributions of avalanche size and life time emerged from the statistics of their numerical calculations, Fig. 22. And, these exponents have been demonstrated to satisfy the standard exponent relations known for SOC processes [128]. There are other statistical verifications of SOC behavior from both the 1D and 2D dynamical modeling results, which we shall not discuss here.

More interestingly, Klimas et al. [69] recently demonstrated that the velocity fluctuations of their 2D simulation of the dynamical model exhibited both SOC and multifractal characteristics very similar to those expected for intermittent MHD turbulence; thus, corroborating with some of the basic concepts of interwoven connection and crossover phenomenon between SOC/FSOC and intermittent turbulence as espoused in the previous section.

Future Directions

We have now come to perhaps the most important part of this article. As discussed in the introduction, the development of the modern concepts of dynamical complexity in space plasmas is very recent. Despite a flurry of activities, current investigations – both from the theoretical and observational points of view – have barely scratched the surface of what may be achieved and understood. For example, we need to know more about the complexity behavior of the plasmas from the kinetic point of view. We also need to know how the plasma particle distributions respond to the fluctuations of its turbulent environment. In addition, the global space environment is generally inhomogeneous with complex initial and boundary conditions. Thus we need to address processes that are statistically nonstationary and spatially inhomogeneous as well as those that evolve temporally and nonuniformly. We also need to understand how the complexity and multifractal phenomena of space plasmas influence the global dynamics of the space plasma environment.

Throughout the discourse, we have placed our emphasis of dynamical complexity in space plasmas on the basic underlying physics. We have come to realize that such fundamental ideas must come from the understanding of the complicated stochastic behavior arisen from the interactions of the multitudes of coherent structures that are prevalent in the natural development of the dynamics of space plasmas. We learned that there were various types of invariant concepts related to coarse-graining, which characterize the primary phenomena that we are attempting to study, analyze and understand. We studied a suite of statistical methodology to analyze complexity, some of which are new and some are carryovers from other fields of complexity, including hydrodynamic turbulence.

We realized that most of the so-called theories that exist today in dynamical complexity of space plasmas are mainly phenomenological concepts and generally mimic those that were prevalent before the modern developments of complexity. We know that these theories cannot carry us too far as we know that they had not advanced the true understanding of complexity phenomena in the past. For this reason, we have broken the tradition in this article and tried not to continually compare our statistically acquired results and theoretically developed understandings with the prevalent classical ideas and numerology. Such comparisons abound in the literature and generally murk the basic concepts that we are trying to convey.

We have learned that most of the phenomena of complexity in space plasmas are scale invariant concepts and yet very little theory has been developed to try to explore these invariant characteristics. We are still stuck with the dogma that we must develop our theories from the "basic elemental equations", while fully aware that the phenomena that we are trying to address generally have nothing to do with these equations directly. After all, nearly every invariant result gained from the coarse-graining concept breaks most of the symmetries and invariance properties of the basic elemental equations.

We have learned that the elemental dissipation such as classical resistivity and viscosity generally has nothing to do with the apparent dissipation that are everywhere in dynamical complexity. And yet every existing theoretical idea insists on including these almost irrelevant dissipative terms in its basic development.

We learned that intermittency generally involve large deviations from the mean and yet we insist on fitting every statistical result into neatly packed curves as if all physics and theories must result in smooth curves in parametric studies. It is time to bring in the ideas of Lebesgue measure, extreme value statistics and thermodynamics of rare events, as well as rank-ordered groupings and associated analyzes to our future developments of new theories and statistical methods

With enthusiasm, we look forward to the dawning of a new era from the future horizon when these new concepts and theoretical understandings finally establish their footings in the modern development of dynamical complexity in space plasmas.

Acknowledgments

The author is extremely grateful to Cheng–chin Wu for his unremitting collaboration and for many of the numerical results that are displayed in this article. He is indebted to Vassilis Angelopoulos, Roberto Bruno, Sandra Chapman, Abe Chian, Giuseppe Consolini, Marius Echim, Bogdan Hnat, Jack Jasperse, Charlie Kennel, Margy Kivelson, Alex Klimas, Hervé Lamy, Tony Lui, Eckart Marsch, Bill Matthaeus, John Podesta, Surja Sharma, Sunny Tam, Vadim Uritsky, Dimitri Vassiliadis, Zoltan Vöröss, Nick Watkins, Jim Weygand, and Gary Zank for cooperative research and/or useful discussions. He is very appreciative to those who kindly allowed him to quote and reprint their published results and to Rich Chang for a critical reading of the manuscript and for his helpful suggestions. The author wishes to thank the Section Editor of Chaos and Complexity in Astrophysics for this encyclopedia volume, for inviting him to contribute this article and for his careful editing and useful suggestions. He is most grateful to members of his family for their constant encouragement and particularly to his wife, Monica, for her understanding and persistent spurring. Permissions from Elsevier. Springer Science and Business Media, the American Astrophysical Society, and the Italian Society of Physics and the liberal copyright policies of the American Physical Society, the American Institute of Physics, the American Geophysical Union, and the European Geophysical Union, which allowed the author to reprint figures and illustrations from their publications, are gratefully acknowledged. This project was partially supported by the AFOSR, NASA and NSF of the US government.

Bibliography

Primary Literature

1. Alexandrova O, Mangeney A, Maksimovic M, Cornilleau-Wehrlin N, Bosqued JM, André M (2006) Alfvén vortex filaments observed in magnetosheath downstream of a quasi-perpendicular bow shock. J Geophys R 111:A12208; doi:10.1029/2006JA011934
2. Angelopoulos V, Mukai T, Kokubun S (1999) Evidence for intermittency in Earth's plasma sheet and implications for self-organized criticality. Phys Plasmas 6:4161–4168
3. Bak P, Tang C, Wiesenfeld K (1987) Self-organized criticality: an explanation of 1/f noise. Phys Rev Lett 59:381–384
4. Baker DN, Klimas AJ, McPherson RL, Büchner J (1990) The evolution from weak to strong geomagnetic activity: An interpretation in terms of deterministic chaos. Geophys Res Lett 17:41–44
5. Balatoni J, Renyi A (1956) Pub Math Inst Hung Acad Sci 1:9

6. Benzi R, Ciliberto S, Tripiccione R, Baudet C, Massaioli F, Succi S (1993) Extended self-similarity in turbulent flows. Phys Rev E 48:R29–32
7. Biskamp D (1993) Nonlinear magnetohydrodynamics. Cambridge University Press, Cambridge
8. Biskamp D (2003) Magnetohydrodynamic turbulence. Cambridge University Press, Cambridge
9. Bruno R, Bavassano B, Pietropaolo E, Carbone V, Veltri P (1999) Effects of intermittency on interplanetary velocity and magnetic field fluctuations anisotropy. Geophys Res Lett 26:3185–3188
10. Bruno R, Carboni V, Veltri P, Pietropaolo E, Bavassano B (2001) Identifying intermittency events in the solar wind. Planet Space Sci 49:1201–1210
11. Bruno R, Carbone V, Chapman S, Hnat B, Noullez A, Sorriso-Valvo L (2007) Intermittent character of interplanetary magnetic field fluctuations. Phys Plasmas 14:032901
12. Burlaga LF (1991) Intermittent turbulence in the solar wind. J Geophys Res 96:5847–5851
13. Burlaga LF (1991) Multifractal structure of the interplanetary magnetic field: Voyager 2 observations near 25 AU, 1987–1988. Geophys Res Lett 18:69–72
14. Burlaga LF (1991) Multifractal structure of speed fluctuations in recurrent streams at 1AU and near 6 AU. Geophys Res Lett 18:1651–1654
15. Burlaga LF (1993) Intermittent turbulence in large-scale velocity fluctuations at 1 AU near solar maximum. J Geophys R 98:17467–17473
16. Carbone V (1993) Cascade model for intermittency in fully developed magneto-hydrodynamic turbulence. Phys Rev Lett 71:1546–1548
17. Carbone V (1994) Scaling exponents of the velocity structure functions in the interplanetary medium. Ann Geophys 12:585–590
18. Carbone V, Bruno R (1996) Cancellation exponents and multifractal scaling laws in the solar wind magnetohydrodynamic turbulence. Ann Geophys 14:777–785
19. Carbone V, Veltri P, Bruno R (1995) Experimental evidence for differences in the extended self-similarity scaling laws between fluid and magnetohydrodynamic turbulent flows. Phys Rev Lett 75:3110–3113
20. Carbone V, Bruno R, Veltri P (1996) Evidence for extended self-similarity in hydromagnetic turbulence. Geophys Res Lett 23:121–124
21. Castaing B, Gagne Y, Hopfinger EJ (1990) Velocity probability density functions of high Reynolds number turbulence. Physica D 46:177–200
22. Chang T (1992) Low-dimensional behavior and symmetry breaking of stochastic systems near criticality – Can these effects be observed in space and in the laboratory? IEEE Trans Plasma Sci 20:691–694
23. Chang T (1998) Sporadic localized reconnections and multiscale intermittent turbulence in the magnetotail. In: Horwitz JL, Gallagher DI, Peterson WK (eds) Geospace mass and energy flow. Geophysical Monograph, vol 104. AGU, Washington, pp 193–199
24. Chang T (1999) Self-organized criticality, multifractal spectra, sporadic localized reconnections and intermittent turbulence in the magnetotail. Phys Plasmas 6:4137–4145
25. Chang T (2001) An example of resonances, coherent structures and topological phase transitions – the origin of the low frequency broadband spectrum in the auroral zone. Nonlinear Process Geophys 8:175–179
26. Chang T (2001) Colloid-like behavior and topological phase transitions in space plasmas: intermittent low frequency turbulence in the auroral zone. Phys Scr T89:80–83
27. Chang T, Wu CC (2007) Dynamical complexity, intermittent turbulence, coarse-grained dissipation, criticality and multifractal processes. In: Shaikh D (ed) Turbulence and nonlinear processes in astrophysical plasmas. AIP Proc 932:161–166
28. Chang T, Wu CC (2008) Rank-ordered multifractal spectrum for intermittent fluctuations. Phys Rev E 77:045401(R)
29. Chang T, Hankey A, Stanley HE (1973) Generalized scaling hypothesis in multicomponent systems, vol 1. Classication of critical points by order and scaling at tricriical points. Phys Rev B 8:346–364
30. Chang T, Nicoll JF, Young JF (1978) A closed-form differential renormalization-group generator for critical dynamics. Phys Lett A 67:287–290
31. Chang T, Vvedensky DD, Nicoll JF (1992) Differential renormalization-group generators for static and dynamic critical phenomena. Phys Reports 217:279–362
32. Chang T, Tam SWY, Wu CC, Consolini G (2003) Complexity, forced and/or self-organized criticality, and topological phase transitions in space plasmas. Space Sci Rev 107:425–445
33. Chang T, Tam SWY, Wu CC (2004) Complexity induced anisotropic bimodal intermittent turbulence in space plasmas. Phys Plasmas 11:1287–1299
34. Chang T, Tam SWY, Wu CC (2006) Complexity in space plasmas – a brief review. Space Sci Rev 122:281–291
35. Chapman SC, Watkins NW, Dendy RO, Helander P, Rowlands G (1998) A simple avalanche model as an analogue for the magnetospheric activity. Geophys Res Lett 25:2397–2400
36. Chapman SC, Dendy RO, Rowlands G (1999) A sandpile model with dual scaling regimes for laboratory, space, and astrophysical plasmas. Phys Plasmas 6:4169–4177
37. Consolini G (1997) Sandpile cellular automata and the magnetospheric dynamics. In: Aiello S, Iucci N, Sironi G, Treves A, Villante U (ed) Cosmic Physics in the Year 2000. Proc VIII GIFCO Conference, Società Italiana di Fisica, Bologna, pp 123–126
38. Consolini G, Chang T (2001) Magnetic field topology and criticality in geotail dynamics: Relevance to substorm phenomena. Space Sci Rev 95:309
39. Consolini G, Chang T (2002) Complexity, magnetic field topology, criticality, and metastability in magnetotail dynamics. J Atmos Solar Terr Phys 64:541
40. Consolini G, Lui ATY (2000) Symmetry Breaking and Nonlinear Wave-Wave Interaction in Current disruption: Possible evidence for a phase Transition. In: Ohtani S et al (eds) Magnetospheric Current Systems. Geophysical Monograph, vol 118. AGU, Washington, p 395
41. Consolini G, Marcucci MF, Candidi M (1996) Multifractal structure of auroral electrojet index data. Phys Rev Lett 76:4082–4085
42. Consolini G, Chang T, Lui ATY (2005) Complexity and topological disorder in the Earth's magnetotail dynamics. In: Sharma AS, Kaw PK (eds) Nonequilibrium phenomena in plasmas. Springer, Dordrecht, pp 51–69
43. Daubechies I (1992) Ten lectures on wavelets. Society for Industrial and Applied Mathematics, Philadelphia, p 357

44. Echim MM, Lamy H, Chang T (2007) Multipoint observations of intermittency in the cusp regions. Nonlinear Processes in Geophysics 14:525–534
45. Einaudi G, Velli M (1999) The distribution of flares, statistics of magnetohydrodynamic turbulence and coronal heating. Phys Plasmas 6:4146–4153
46. Farge M (1992) Wavelet transforms and their applications to turbulence. Ann Rev Fluid Mech 24:395–457
47. Feynman J, Ruzmaikin A (1994) Distributions of the interplanetary field revisited. J Geophys Res 99:17645–17651
48. Forman M, Burlaga LF (2003) Exploring the Castaing distribution function to study intermittence in the solar wind at L1 in June 2000. In: Velli M, Bruno R, Malara F (eds) Solar wind ten. AIP Conf Proc 679:554–557
49. Forster D, Nelson DR, Stephen MJ (1977) Large distance and long time properties of a randomly stirred fluid. Phys Rev A 16:732–749
50. Grassberger P, Procaccia I (1983) Characterization of strange attractors. Phys Rev Lett 50:346–349
51. Grossmann S, Lohse D, Reeh A (1997) Application of extended self-similarity in turbulence. Phys Rev E 56:5473–5478
52. Haar A (1910) Zur Theorie der Orthogonalen Functionensysteme. Math Ann 69:331–371
53. Halperin BI, Hohenberg PC, Ma SK (1972) Calculation of dynamical critical properties using Wilson's expansion methods. Phys Rev Lett 29:1548–1551
54. Halsey T, Jensen MH, Kadanoff LP, Procaccia I, Schraiman BI (1986) Fractal measures and their singularities: The characterization of strange sets. Phys Rev A 33:1141–1151
55. Hentschel HGE, Procaccia I (1983) The infinite number of generalized dimensions of fractals and strange attractors. Physica 8D:435–444
56. Hnat B, Chapman SC, Rowlands G, Watkins NW, Farrell WM (2002) Finite size scaling in the solar wind magnetic field energy density as seen by WIND. Geophys Res Lett 29:1446; doi:10.1029/2001GL014587
57. Hnat B, Chapman SC, Rowlands G, Watkins NW, Freeman MP (2003) Scaling in long term data sets of geomagnetic indices and solar wind ε as seen by WIND spacecraft. Geophys Res Lett 30:2174; doi:10.1029/2003GL018209
58. Horbury TA, Balogh A, Forsyth RJ, Smith EJ (1995) Ulysses magnetic field observations of fluctuations within polar coronal flows. Ann Geophys 13:105–107
59. Horbury TA, Balogh A, Forsyth RJ, Smith EJ (1995) Observations of evolving turbulence in the polar solar wind. Geophys Res Lett 22:3401–3404
60. Horbury TA, Balogh A, Forsyth RJ, Smith EJ (1997) Ulysses observations of intermittent heliospheric turbulence. Adv Space Phys 19:847–850
61. Hosino M, Nishida A, Yamamoto T, Kokubun S (1994) Turbulent magnetic field in the distant magnetotail: Bottom-up process of plasmoid formation? Geophys Res Lett 21:2935–2938
62. Hwa T, Kardar M (1992) Avalanches, hydrodynamics, and discharge events in models of sandpiles. Phys Rev A 45:7002–7023
63. Jensen HJ (1998) Self-organized criticality – Emergent complex behavior in physical and biological systems. Cambridge University Press, Cambridge
64. Jiang GS, Wu CC (1999) A high-order WENO finite difference scheme for the equations of ideal magnetohydrodynamics. J Comp Phys 150:561–594
65. Klimas AJ, Baker DN, Roberts DA, Fairfield DH, Büchner J (1992) A nonlinear dynamical analogue model of geomagnetic activity. J Geophys Res 97:12253–12266
66. Klimas AJ, Valdivia JA, Vassiliadis D, Baker DN, Hesse M, Takalo J (2000) Self-organized criticality in the substorm phenomenon and its relation to localized reconnection in the magnetospheric plasma sheet. J Geophys Res 105:18765–18780
67. Klimas AJ, Uritsky VM, Vassiliadis D, Baker DN (2004) Reconnection and scale-free avalanching in a driven current-sheet model. J Geophys Res 109:A02218; doi:10.1029/2003JA010036
68. Klimas AJ, Uritsky VM, Vassiliadis D, Baker DN (2005) A mechanism for the loading-unloading substorm cycle missing in MHD global magnetospheric simulation models. Geophys Res Lett 32:L14108; doi:10.1029/2005GL022916
69. Klimas AJ, Uritsky VM, Paczuski M (2007) Self-organized criticality and intermittent turbulence in an MHD current sheet with a threshold Instability. arXiv:astro-ph/0701486v2
70. Langmuir I (1928) Oscillations in ionized gases. Proc Natl Acad Sci 14:628
71. Leubner LP (2004) Fundamental issues on kappa-distributions in space plasmas and interplanetary proton distributions. Phys Plasmas 11:1308–1316
72. Leubner MP, Vörös Z (2005) A nonextensive entropy approach to solar wind intermittency. Astrophys J 618:547–555
73. Lu ET (1995) Avalanches in continuum driven dissipative systems. Phys Rev Lett 74:2511–2514
74. Lu ET, Hamilton RJ (1991) Avalanches and the distribution of solar flares. Astrophys J 380:L89–L92
75. Lui ATY (1996) Current disruption in the Earth's magnetosphere: observations and models. J Geophys Res 101:13067–13088
76. Lui ATY (2001) Multifractal and intermittent nature of substorm-associated magnetic turbulence in the magnetotail. J Atmos Solar Terr Phys 63:1379–1385
77. Lui ATY (2002) Multiscale phenomena in the near-Earth magnetosphere. J Atmos Solar Terr Phys 64:125
78. Lui ATY, Lopez RE, Krimigis SM, McEntire RW, Zanetti LJ, Potemra TA (1988) A case study of magnetotail current sheet disruption and diversion. Geophys Res Lett 15:721–724
79. Lui ATY, Chang CL, Mankofsky A, Wong HK, Winske D (1991) A cross-field current instability for substorm expansions. J Geophys Res 96:11389–11401
80. Lui ATY, Liou K, Newell PT, Meng CI, Ohtani SI, Kokubun S, Ogino T, Brittnacher M, Parks G (1998) Plasma and magnetic flux transport associated with auroral breakups. Geophys Res Lett 25:4059–4062
81. Lui ATY, Chapman SC, Liou K, Newell PT, Meng CI, Brittnacher M, Parks GK (2000) Is the dynamic magnetosphere an avalanching system? Geophys Res Lett 27:911–914
82. Lui ATY, Zheng Y, Zhang Y, Rème H, Dunlop MW, Gustafsson G, Mende SB, Mouikis C, Kistler LM (2006) Cluster observation of plasma flow reversal in the magnetotail during a substorm. Ann Geophys 24:2005–2013
83. Mandelbrot B (1977) Fractals: Form, chance and dimension. Freeman, San Francisco

84. Marsch E, Liu S (1993) Structure functions and intermittency of velocity fluctuations in the inner solar wind. Ann Geophys 11:227–238
85. Marsch E, Tu CY (1994) Non-Gaussian probability distributions of solar wind fluctuations. Ann Geophys 12:1127–1138
86. Marsch E, Tu CY (1997) Intermittency, non-Gaussian statistics and fractal scaling of MHD fluctuations in the solar wind. Nonlinear Process Geophys 4:101–124
87. Marsch E, Tu CY, Rosenbauer H (1996) Multifractal scaling of the kinetic energy flux in solar wind turbulence. Ann Geophys 14:259–269
88. Matthaeus WH, Goldstein ML (1982) Measurements of the rugged invariants of magnetohydrodynamic turbolence in the solar wind. J Geophys Res 87:6011–6028
89. Matthaeus WH, Goldstein ML, Roberts DA (1990) Evidence for the presence of quasi-two-dimensional nearly incompressible fluctuations in the solar wind. J Geophys Res 95:20673–20683
90. Matthaeus WH, Qin G, Bieber JW, Zank GP (2003) Nonlinear collisionless perpendicular diffusion of charged particles. Astrophys J Lett 590:L53–L56
91. Meneveau C (1991) Analysis of turbulence in the orthonormal wavelet representation. J Fluid Mech 232:469–521
92. Meneveau C, Sreenivasan KR (1987) Simple multifractal cascade model for fully developed turbulence. Phys Rev Lett 59:1424–1427
93. Milovanov AV, Zelenyi LM, Zimbardo G (1996) Fractal structures and power law spectra in the distant Earth's magnetotail. J Geophys Res 101:19903–19910
94. Morlet J (1983) Sampling theory and wave propagation. In: Chen CH (ed) Issues on acoustic signal/image processing and recognition, NATO ASI, vol F1. Springer, Berlin
95. Müller WC, Biskamp D (2002) Scaling properties of three-dimensional magnetohydrodynamic turbulence. Phys Rev Lett 84:475–478
96. Ott E (1993) Chaos in dynamical systems. Cambridge University Press, New York
97. Pagel C, Balogh A (2001) A study of magnetic fluctuations and their anomalous scaling in the solar wind: the Ulysses fast-latitude scan. Nonlinear Process Geophys 8:313–330
98. Parisi G, Frisch U (1985) On the singularity structure of fully developed turbulence. In: Ghil M, Benzi R, Parisi G (eds) Turbulence and predictability in geophysical fluid dynamics. Proc Intern School of Physics, 'Erico Fermi' 1983, Varenna, Italy. North Holland, Amsterdam, pp 84–87
99. Pinheiro MJ (2007) Plasmas: the genesis of the word. arXiv:physics/0703260v1
100. Politano H, Pouquet A, Carbone V (1998) Determination of anomalous exponents of structure functions in two-dimensional magnetohydrodynamic turbulence. Europhys Lett 43:516–521
101. Roberts DA, Baker DN, Klimas AJ, Bargatze LF (1991) Indications of low dimensionality in magnetospheric dynamics. Geophys Res Lett 18:151–154
102. Rusmaikin A, Lyannaya IR, Styashkin VA, Yeroshenko E (1993) The spectrum of the interplanetary magnetic field near 1.3 AU. J Geophys Res 98:13303–13306
103. Shan LH, Hansen P, Goertz CK, Smith RA (1991) Chaotic appearance of the AE index. Geophys Res Lett 18:147–150
104. Shan LH, Goertz CK, Smith RA (1991) On the embedding-dimension analysis of AE and AL time series. Geophys Res Lett 18:1647–1650
105. Sharma AS, Vassiliadis DV, Papadopoulos K (1993) Reconstruction of low-dimensional magnetospheric dynamics by singular spectrum analysis. Geophys Res Lett 20:335–338
106. Sitnov GL, Sharma AS, Papadopoulos K, Vassiliadis D, Valdivia JA, Klimas AJ, Baker DN (2000) Phase transition-like behavior of the magnetosphere during substorms. J Geophys Res 105:12955–12974
107. Sorriso-Valvo L, Carbone V, Veltri P, Consolini G, Bruno R (1999) Intermittency in the solar wind turbulence through probability distribution functions of fluctuations. Geophys Res Lett 26:1801–1804
108. Sorriso-Valvo L, Carbone V, Giuliani P, Veltri P, Bruno R, Antoni V, Martines E (2001) Intermittency in plasma turbulence. Planet Space Sci 49:1193–1200
109. Sundkvist D, Krasnoselskikh V, Shukla PK, Vaivads A, André M, Buchert S, Rème H (2005) In situ multi-satellite detection of coherent vortices as a manifestation of Alfvénic turbulence. Nature 436:825–828
110. Takalo J, Timonen J, Klimas AJ, Valdivia JA, Vassiliadis D (2001) A coupled map as a model of the dynamics of the magnetotail current sheet. J Atmos Solar Terr Phys 63:1407–1414
111. Tam SWY, Chang T (2005) Energization of ions by bimodal intermittent fluctuations. In: Lui ATY, Kamide Y, Consolini G (eds) Multiscale coupling of sun-earth processes. Elsevier, Netherlands, pp 375
112. Tam SWY, Chang T, Kintner PM, Klatt E (2005) Intermittency analyses on the SIERRA measurements of the electric field fluctuations in the auroral zone. Geophys Res Lett 32:L05109; doi:10.1029/2004GL021445
113. Taylor GI (1938) The spectrum of turbulence. Proc R Soc Lond A164:476–490
114. Taylor JB (1974) Relaxation of toroidal plasma and generation of reverse magnetic fields. Phys Rev Lett 33:1139–1141
115. Tetreault D (1991) Theory of electric fields in the auroral acceleration region. J Geophys Res 96:3549–3563
116. Tetreault D (1992) Turbulent relaxation of magnetic fields, 1. Coarse-grained dissipation and reconnection. J Geophys Res 97:8531–8540
117. Tetreault D (1992) Turbulent relaxation of magnetic fields, 2. Self-organization and intermittency. J Geophys Res 97:8541–8547
118. Tonks L (1967) Am J Phys 35:857–858
119. Tonks L, Langmuir I (1929) Oscillations in ionized gases. Phys Rev 33:195–210
120. Tu CY, Marsch E (1995) MHD structures, waves and turbulence in the solar wind – observations and theories. Kluwer, Dordrecht
121. Tu CY, Marsch E, Rosenbauer H (1996) An extended structure function model and its application to the analysis of solar wind intermittency properties. Ann Geophys 14:270–285
122. Uritsky VM, Klimas AJ (2005) Hysteresis-controlled instability waves in a scale-free driven current sheet model. Nonlinear Process Geophys 12:827–833
123. Uritsky VM, Klimas AJ, Valdivia JA, Vassiliadis D, Baker DN (2001) Stable critical behavior and fast field annihilation in a magnetic field reversal model. J Atmos Solar Terr Phys 63:1425–1433

124. Uritsky VM, Klimas AJ, Vassiliadis D, Chua D, Parks G (2002) Scale-free statistics of spatiotemporal auroral emissions as depicted by POLAR UVI images: Dynamic magnetosphere is an avalanching system. J Geophys Res 107:1426; doi:10.1029/2001JA000281
125. Uritsky VM, Klimas AJ, Vassiliadis D (2002) Multiscale dynamics and robust critical scaling in a continuum current sheet model. Phys Rev E 65:046113
126. Uritsky VM, Paczuski M, Davila JM, Jones SI (2007) Coexistence of self-organized criticality and intermittent turbulence in the solar corona. Phys Rev Lett 99:025001
127. Vahnstein S, Sreenivasan K, Pierrehumbert R, Kashyap V, Juneja A (1994) Scaling exponents for turbulence and other random processes and their relationships with multifractal structure. Phys Rev E 50:1823–1835
128. Vespignani A, Zapperi S (1998) How self-organized criticality works: A unified mean-field picture. Phys Rev E 57:6345–6362
129. Vörös Z (2000) On multifractality of high-latitude geomagnetic fluctuations. Ann Geophys 18:1273–1283
130. Vörös Z, Baumjohann W, Nakamura R, Runov A, Zhang TL, Volwerk M, Eichelberger HU, Balogh A, Horbury TS, Glassmeier KH, Klecker B, Rème H (2003) Multi-scale magnetic field intermittence in the plasma sheet. Ann Geophys 21:1955–1964
131. Vörös Z, Baumjohann W, Nakamura R, Runov A, Volwerk M, Zhang TL, Balogh A (2004) Wavelet analysis of magnetic turbulence in the Earth's plasma sheet. Phys Plasmas 11:1333–1338
132. Vörös Z, Baumjohann W, Nakamura R, Runov A, Volwerk M, Schwarzl H, Balogh A, Rème H (2005) Dissipation scales in the Earth's plasma sheet estimated from cluster measurements. Nonlinear Process Geophys 12:725–732
133. Watkins NW, Chapman SC, Dendy RO, Rowlands G (1999) Robustness of collective behaviour in a strongly driven avalanche model: magnetospheric implications. Geophys Res Lett 26:2617–2620
134. Weygand JM, Kivelson MG, Khurana KK, Schwarzl HK, Thompson SM, McPherron RL, Balogh A, Kistler LM, Goldstein ML, Borovsky J, Roberts DA (2005) Plasma sheet turbulence observed by Cluster, vol II. J Geophys Res 110:A01205; doi:10.1029/2004JA010581
135. Wilson KG, Kogut J (1994) The renormalization group and the ε expansion. Phys Reports 12C:75–200
136. Wu CC, Chang T (2000) 2D MHD simulation of the emergence and merging of coherent structures. Geophys Res Lett 27:863–866
137. Wu CC, Chang T (2001) Further study of the dynamics of two-dimensional MHD coherent structures – a large scale simulation. J Atmos Solar Terr Phys 63:1447–1453
138. Wu CC, Chang T (2005) Intermittent turbulence in 2D MHD simulation. In: Lui ATY, Kamide Y, Consolini G (eds) Multiscale coupling of sun-earth processes. Elsevier, Netherlands, p 321
139. Yordanova E, Grezesiak M, Wernik AW, Popielawska B, Stasiewicz K (2005) Multifractal structure of turbulence in the magnetospheric cusp. Ann Geophys 22:2431–2440

Books and Reviews

Bruno R, Carbone V (2005) The solar wind as a turbulence laboratory. Living Rev Solar Phys 2:4; http://www.livingreviews.org/lrsp-2005-4. Accessed 15 June 2007

Chang T, Chapman S, Klimas A (eds) (2001) Forced and/or self-organized criticality (FSOC) in space plasmas. J Atmos Solar Terr 63:1359–1453

Forster D (1975) Hydrodynamic fluctuations, broken symmetry, and correlation functions. Benjamin, Reading

Frisch U (1995) Turbulence. Cambridge University Press, Cambridge

Lavenda BH (1995) Theromodynamics of Extremes. Albion Publishers, Chichester

Leubner MP, Baumjohann W, Chian ACL (eds) (2006) Advances in space environment research, vol 2. Space Sci Rev 122:1–337

Lui ATY, Kamide Y, Consolini G (eds) (2005) Multiscale coupling of sun-earth processes. Elsevier, the Netherlands

Ma SK (1976) Modern theory of critical phenomena. Benjamin, Reading

Sharma AS, Kaw PK (eds) (2005) Nonequilibrium phenomena in plasmas. Astrophysics and Space Science Library, vol 321. Springer, Dordrecht, pp 1–340

Sornette D (2000) Critical phenomena in natural sciences; Chaos, fractals, selforganization and disorder: concepts and tools. Springer, Berlin

Spectral Theory of Dynamical Systems

Mariusz Lemańczyk
Faculty of Mathematics and Computer Science,
Nicolaus Copernicus University, Toruń, Poland

Article Outline

Glossary

Spectral decomposition of a unitary representation If $\mathcal{U} = (U_a)_{a\in\mathbb{A}}$ is a continuous unitary representation of a locally compact second countable (l.c.s.c.)

Abelian group $\mathbb{A}$ in a separable Hilbert space H then a decomposition $H = \bigoplus_{i=1}^{\infty} \mathbb{A}(x_i)$ is called *spectral* if $\sigma_{x_1} \gg \sigma_{x_2} \gg \dots$ (such a sequence of measures is also called *spectral*); here $\mathbb{A}(x) := \overline{span}\{U_a x \colon a \in \mathbb{A}\}$ is called the *cyclic space* generated by $x \in H$ and σ_x stands for the spectral measure of x.

Maximal spectral type and the multiplicity function of $\mathcal{U}$ The *maximal spectral type* $\sigma_{\mathcal{U}}$ of $\mathcal{U}$ is the type of σ_{x_1} in any spectral decomposition of H; the *multiplicity function* $M_{\mathcal{U}} \colon \widehat{\mathbb{A}} \to \{1, 2, \dots\} \cup \{+\infty\}$ is defined $\sigma_{\mathcal{U}}$-a.e. and $M_{\mathcal{U}}(\chi) = \sum_{i=1}^{\infty} 1_{Y_i}(\chi)$, where $Y_1 = \widehat{\mathbb{A}}$ and $Y_i = \operatorname{supp} \mathrm{d}\sigma_{x_i}/\mathrm{d}\sigma_{x_1}$ for $i \geq 2$.

A representation $\mathcal{U}$ is said to have *simple spectrum* if H is reduced to a single cyclic space. The multiplicity is *uniform* if there is only one essential value of $M_{\mathcal{U}}$. The essential supremum of $M_{\mathcal{U}}$ is called the *maximal spectral multiplicity*. $\mathcal{U}$ is said to have *discrete spectrum* if H has an orthonormal base consisting of eigenvectors of $\mathcal{U}$; $\mathcal{U}$ has *singular* (*Haar*, *absolutely continuous*) spectrum if the maximal spectral type of $\mathcal{U}$ is singular with respect to (equivalent to, absolutely continuous with respect to) a Haar measure of $\widehat{\mathbb{A}}$.

Koopman representation of a dynamical system $\mathcal{T}$ Let $\mathbb{A}$ be a l.c.s.c. (and not compact) Abelian group and $\mathcal{T} \colon a \mapsto T_a$ a representation of $\mathbb{A}$ in the group $Aut(X, \mathcal{B}, \mu)$ of (measure-preserving) automorphisms of a standard probability Borel space $(X, \mathcal{B}, \mu)$. The *Koopman representation* $\mathcal{U} = \mathcal{U}_{\mathcal{T}}$ of $\mathcal{T}$ in $L^2(X, \mathcal{B}, \mu)$ is defined as the unitary representation $a \mapsto U_{T_a} \in U(L^2(X, \mathcal{B}, \mu))$, where $U_{T_a}(f) = f \circ T_a$.

Ergodicity, weak mixing, mild mixing, mixing and rigidity of $\mathcal{T}$ A measure-preserving $\mathbb{A}$-action $\mathcal{T} = (T_a)_{a \in \mathbb{A}}$ is called *ergodic* if $\chi_0 \equiv 1 \in \widehat{\mathbb{A}}$ is a simple eigenvalue of $\mathcal{U}_{\mathcal{T}}$. It is *weakly mixing* if $\mathcal{U}_{\mathcal{T}}$ has a continuous spectrum on the subspace $L_0^2(X, \mathcal{B}, \mu)$ of zero mean functions. $\mathcal{T}$ is said to be *rigid* if there is a sequence (a_n) going to infinity in $\mathbb{A}$ such that the sequence $(U_{T_{a_n}})$ goes to the identity in the strong (or weak) operator topology; $\mathcal{T}$ is said to be *mildly mixing* if it has no non-trivial rigid factors. We say that $\mathcal{T}$ is *mixing* if the operator equal to zero is the only limit point of $\{U_{T_a}|_{L_0^2(X, \mathcal{B}, \mu)} \colon a \in \mathbb{A}\}$ in the weak operator topology.

Spectral disjointness Two $\mathbb{A}$-actions S and $\mathcal{T}$ are called *spectrally disjoint* if the maximal spectral types of their Koopman representations $\mathcal{U}_{\mathcal{T}}$ and $\mathcal{U}_S$ on the corresponding L_0^2-spaces are mutually singular.

SCS property We say that a Borel measure σ on $\widehat{\mathbb{A}}$ satisfies the *strong convolution singularity* property (SCS property) if, for each $n \geq 1$, in the disintegration (given by the map $(\chi_1, \dots, \chi_n) \mapsto \chi_1 \cdot \ldots \cdot \chi_n$) $\sigma^{\otimes n} = \int_{\widehat{\mathbb{A}}} \nu_\chi \, d\sigma^{(n)}(\chi)$ the conditional measures ν_χ are atomic with exactly $n!$ atoms ($\sigma^{(n)}$ stands for the nth convolution $\sigma * \ldots * \sigma$). An $\mathbb{A}$-action $\mathcal{T}$ satisfies the SCS property if the maximal spectral type of $\mathcal{U}_{\mathcal{T}}$ on L_0^2 is a type of an SCS measure.

Kolmogorov group property An $\mathbb{A}$-action $\mathcal{T}$ satisfies the *Kolmogorov group property* if $\sigma_{\mathcal{U}_{\mathcal{T}}} * \sigma_{\mathcal{U}_{\mathcal{T}}} \ll \sigma_{\mathcal{U}_{\mathcal{T}}}$.

Weighted operator Let T be an ergodic automorphism of $(X, \mathcal{B}, \mu)$ and $\xi \colon X \to \mathbb{T}$ be a measurable function. The (unitary) operator $V = V_{\xi,T}$ acting on $L^2(X, \mathcal{B}, \mu)$ by the formula $V_{\xi,T}(f)(x) = \xi(x) f(Tx)$ is called a *weighted operator*.

Induced automorphism Assume that T is an automorphism of a standard probability Borel space $(X, \mathcal{B}, \mu)$. Let $A \in \mathcal{B}$, $\mu(A) > 0$. The *induced automorphism* T_A is defined on the conditional space $(A, \mathcal{B}_A, \mu_A)$, where $\mathcal{B}_A$ is the trace of $\mathcal{B}$ on A, $\mu_A(B) = \mu(B)/\mu(A)$ for $B \in \mathcal{B}_A$ and $T_A(x) = T^{k_A(x)} x$, where $k_A(x)$ is the smallest $k \geq 1$ for which $T^k x \in A$.

AT property of an automorphism An automorphism T of a standard probability Borel space $(X, \mathcal{B}, \mu)$ is called *approximatively transitive* (AT for short) if for every $\varepsilon > 0$ and every finite set $f_1, \dots, f_n$ of non-negative L^1-functions on $(X, \mathcal{B}, \mu)$ we can find $f \in L^1(X, \mathcal{B}, \mu)$ also non-negative such that $\| f_i - \sum_j \alpha_{ij} f \circ T^{n_j} \|_{L_1} < \varepsilon$ for all $i = 1, \dots, n$ (for some $\alpha_{ij} \geq 0$, $n_j \in \mathbb{N}$).

Cocycles and group extensions If T is an ergodic automorphism, G is a l.c.s.c. Abelian group and $\varphi \colon X \to G$ is measurable then the pair (T, φ) generates a *cocycle* $\varphi^{(\cdot)}(\cdot) \colon \mathbb{Z} \times X \to G$, where

$$\varphi^{(n)}(x) = \begin{cases} \varphi(x) + \ldots + \varphi(T^{n-1} x) & \text{for} \quad n > 0\,, \\ 0 & \text{for} \quad n = 0\,, \\ -(\varphi(T^n x) + \ldots + \varphi(T^{-1} x)) & \text{for} \quad n < 0\,. \end{cases}$$

(That is $(\varphi^{(n)})$ is a standard 1-cocycle in the algebraic sense for the $\mathbb{Z}$-action $n(f) = f \circ T^n$ on the group of measurable functions on X with values in G; hence the function $\varphi \colon X \to G$ itself is often called a *cocycle*.)

Assume additionally that G is compact. Using the cocycle φ we define a *group extension* T_φ on $(X \times G, \mathcal{B} \otimes \mathcal{B}(G), \mu \otimes \lambda_G)$ (λ_G stands for Haar measure of G), where $T_\varphi(x, g) = (Tx, \varphi(x) + g)$.

Special flow Given an ergodic automorphism T on a standard probability Borel space $(X, \mathcal{B}, \mu)$ and a positive integrable function $f \colon X \to \mathbb{R}^+$ we put

$$X^f = \{(x, t) \in X \times \mathbb{R} \colon 0 \leq t < f(x)\}\,,$$

$$\mathcal{B}^f = \mathcal{B} \otimes \mathcal{B}(\mathbb{R})|_{X^f}\,,$$

and define μ^f as normalized $\mu \otimes \lambda_{\mathbb{R}}|_{X^f}$. By a *special flow* we mean the $\mathbb{R}$-action $T^f = (T^f_t)_{t\in\mathbb{R}}$ under which a point $(x, s) \in X^f$ moves vertically with the unit speed, and once it reaches the graph of f, it is identified with $(Tx, 0)$.

Markov operator A linear operator $J\colon L^2(X, \mathcal{B}, \mu) \to L^2(Y, C, \nu)$ is called *Markov* if it sends non-negative functions to non-negative functions and $J1 = J^*1 = 1$.

Unitary actions on Fock spaces If H is a separable Hilbert space then by $H^{\odot n}$ we denote the subspace of n-tensors of $H^{\otimes n}$ symmetric under all permutations of coordinates, $n \geq 1$; then the Hilbert space $F(H) := \bigoplus_{n=0}^{\infty} H^{\odot n}$ is called a *symmetric Fock space*. If $V \in U(H)$ then $F(V) := \bigoplus_{n=0}^{\infty} V^{\odot n} \in U(F(H))$ where $V^{\odot n} = V^{\otimes n}|H^{\odot n}$.

Definition of the Subject

Spectral theory of dynamical systems is a study of special unitary representations, called Koopman representations (see the glossary). Invariants of such representations are called spectral invariants of measure-preserving systems. Together with the entropy, they constitute the most important invariants used in the study of measure-theoretic intrinsic properties and classification problems of dynamical systems as well as in applications. Spectral theory was originated by von Neumann, Halmos and Koopman in the 1930s. In this article we will focus on recent progresses in the spectral theory of finite measure-preserving dynamical systems.

Introduction

Throughout $\mathbb{A}$ denotes a non-compact l.c.s.c. Abelian group ($\mathbb{A}$ will be most often $\mathbb{Z}$ or $\mathbb{R}$). The assumption of second countability implies that $\mathbb{A}$ is metrizable, σ-compact and the space $C_0(\mathbb{A})$ is separable. Moreover the dual group $\widehat{\mathbb{A}}$ is also l.c.s.c. Abelian.

General Unitary Representations

We are interested in *unitary*, that is with values in the unitary group $U(H)$ of a Hilbert space H, (weakly) continuous representations $V\colon \mathbb{A} \ni a \mapsto V_a \in U(H)$ of such groups (the scalar valued maps $a \mapsto \langle V_a x, y\rangle$ are continuous for each $x, y \in H$).

Let $H = L^2(\widehat{\mathbb{A}}, \mathcal{B}(\widehat{\mathbb{A}}), \mu)$, where $\mathcal{B}(\widehat{\mathbb{A}})$ stands for the σ-algebra of Borel sets of $\widehat{\mathbb{A}}$ and $\mu \in M^+(\widehat{\mathbb{A}})$ (whenever X is a l.c.s.c. space, by $M(X)$ we denote the set of complex Borel measures on X, while $M^+(X)$ stands for the subset of positive (finite) measures). Given $a \in \mathbb{A}$, for $f \in L^2(\widehat{\mathbb{A}}, \mathcal{B}(\widehat{\mathbb{A}}), \mu)$ put

$$V^\mu_a(f)(\chi) = i(a)(\chi) \cdot f(\chi) = \chi(a) \cdot f(\chi) \quad (\chi \in \widehat{\mathbb{A}}),$$

where $i\colon \mathbb{A} \to \widehat{\widehat{\mathbb{A}}}$ is the canonical Pontriagin isomorphism of $\mathbb{A}$ with its second dual. Then $V^\mu = (V^\mu_a)_{a\in\mathbb{A}}$ is a unitary representation of $\mathbb{A}$. Since $C_0(\widehat{\mathbb{A}})$ is dense in $L^2(\widehat{\mathbb{A}}, \mu)$, the latter space is separable. Therefore also direct sums $\bigoplus_{i=1}^{\infty} V^{\mu_i}$ of such type representations will be unitary representations of $\mathbb{A}$ in separable Hilbert spaces.

Lemma 1 (Wiener Lemma) *If $F \subset L^2(\widehat{\mathbb{A}}, \mu)$ is a closed V^μ_a-invariant subspace for all $a \in \mathbb{A}$ then $F = 1_Y L^2(\widehat{\mathbb{A}}, \mathcal{B}(\widehat{\mathbb{A}}), \mu)$ for some Borel subset $Y \subset \widehat{\mathbb{A}}$.*

Notice however that since $\mathbb{A}$ is not compact (equivalently, $\widehat{\mathbb{A}}$ is not discrete), we can find μ continuous and therefore V^μ has no irreducible (non-zero) subrepresentation. From now on only unitary representations of $\mathbb{A}$ in separable Hilbert spaces will be considered and we will show how to classify them.

A function $r\colon \mathbb{A} \to \mathbb{C}$ is called *positive definite* if

$$\sum_{n,m=0}^{N} r(a_n - a_m) z_n \overline{z_m} \geq 0 \tag{1}$$

for each $N > 0$, $(a_n) \subset \mathbb{A}$ and $(z_n) \subset \mathbb{C}$. The central result about positive definite functions is the following theorem (see e. g. [173]).

Theorem 1 (Bochner–Herglotz) *Let $r\colon \mathbb{A} \to \mathbb{C}$ be continuous. Then r is positive definite if and only if there exists (a unique) $\sigma \in M^+(\widehat{\mathbb{A}})$ such that*

$$r(a) = \int_{\widehat{\mathbb{A}}} \chi(a)\, d\sigma(\chi) \quad \textit{for each} \quad a \in \mathbb{A}.$$

If now $\mathcal{U} = (U_a)_{a\in\mathbb{A}}$ is a representation of $\mathbb{A}$ in H then for each $x \in H$ the function $r(a) := \langle U_a x, x\rangle$ is continuous and satisfies (1), so it is positive definite. By the Bochner–Herglotz Theorem there exists a unique measure $\sigma_{\mathcal{U},x} = \sigma_x \in M^+(\widehat{\mathbb{A}})$ (called the *Spectral measure* of x) such that

$$\widehat{\sigma}_x(a) := \int_{\widehat{\mathbb{A}}} i(a)(\chi)\, d\sigma_x(\chi) = \langle U_a x, x\rangle$$

for each $a \in \mathbb{A}$. Since the partial map $U_a x \mapsto i(a) \in L^2(\widehat{\mathbb{A}}, \sigma_x)$ is isometric and equivariant, there exists a unique extension of it to a unitary operator

$W\colon \mathbb{A}(x) \to L^2(\widehat{\mathbb{A}}, \sigma_x)$ giving rise to an isomorphism of $\mathcal{U}|_{\mathbb{A}(x)}$ and V^{σ_x}. Then the existence of a spectral decomposition is proved by making use of separability and a choice of maximal cyclic spaces at every step of an induction procedure. Moreover, a spectral decomposition is unique in the following sense.

Theorem 2 (Spectral Theorem) *If* $H = \bigoplus_{i=1}^{\infty} \mathbb{A}(x_i) = \bigoplus_{i=1}^{\infty} \mathbb{A}(x_i')$ *are two spectral decompositions of* H *then* $\sigma_{x_i} \equiv \sigma_{x_i'}$ *for each* $i \geq 1$.

It follows that the representation $\mathcal{U}$ is entirely determined by types (the sets of equivalent measures to a given one) of a decreasing sequence of measures or, equivalently, $\mathcal{U}$ is determined by its maximal spectral type $\sigma_{\mathcal{U}}$ and its multiplicity function $M_{\mathcal{U}}$.

Notice that eigenvalues of $\mathcal{U}$ correspond to Dirac measures: $\chi \in \widehat{\mathbb{A}}$ *is an eigenvalue* (i. e. *for some* $\|x\| = 1$, $U_a(x) = \chi(a)x$ *for each* $a \in \mathbb{A}$) *if and only if* $\sigma_{\mathcal{U},x} = \delta_\chi$. Therefore $\mathcal{U}$ has a discrete spectrum if and only if the maximal spectral type of $\mathcal{U}$ is a discrete measure.

We refer the reader to [64,105,127,147,156] for presentations of spectral theory needed in the theory of dynamical systems – such presentations are usually given for $\mathbb{A} = \mathbb{Z}$ but once we have the Bochner–Herglotz Theorem and the Wiener Lemma, their extensions to the general case are straightforward.

Koopman Representations

We will consider *measure-preserving* representations of $\mathbb{A}$. It means that we fix a probability standard Borel space $(X, \mathcal{B}, \mu)$ and by $Aut(X, \mathcal{B}, \mu)$ we denote the group of automorphisms of this space, that is $T \in Aut(X, \mathcal{B}, \mu)$ if $T\colon X \to X$ is a bimeasurable (a.e.) bijection satisfying $\mu(A) = \mu(TA) = \mu(T^{-1}A)$ for each $A \in \mathcal{B}$. Consider then a representation of $\mathbb{A}$ in $Aut(X, \mathcal{B}, \mu)$ that is a group homomorphism $a \mapsto T_a \in Aut(X, \mathcal{B}, \mu)$; we write $\mathcal{T} = (T_a)_{a \in \mathbb{A}}$. Moreover, we require that the associated Koopman representation $\mathcal{U}_{\mathcal{T}}$ is continuous. Unless explicitly stated, $\mathbb{A}$-actions are assumed to be *free*, that is we assume that for μ-a.e. $x \in X$ the map $a \mapsto T_a x$ is $1-1$. In fact, since constant functions are obviously invariant for U_{T_a}, that is the trivial character 1 is always an eigenvalue of $\mathcal{U}_{\mathcal{T}}$, the Koopman representation is considered only on the subspace $L_0^2(X, \mathcal{B}, \mu)$ of zero mean functions. We will restrict our attention only to *ergodic* dynamical systems (see the glossary). It is easy to see that $\mathcal{T}$ is ergodic if and only if whenever $A \in \mathcal{B}$ and $A = T_a(A)$ (μ-a.e.) for all $a \in A$ then $\mu(A)$ equals 0 or 1. In case of ergodic Koopman representations, all eigenvalues are simple. In particular, (ergodic) Koopman representations with discrete spectra have simple spectra. The reader is referred to monographs mentioned above as well as to [26,158,177,196,204] for basic facts on the theory of dynamical systems.

The passage $\mathcal{T} \mapsto \mathcal{U}_{\mathcal{T}}$ can be seen as functorial (contravariant). In particular a measure-theoretic isomorphism of $\mathbb{A}$-systems $\mathcal{T}$ and $\mathcal{T}'$ implies spectral isomorphism of the corresponding Koopman representations; hence spectral properties are measure-theoretic invariants. Since unitary representations are completely classified, one of the main questions in the spectral theory of dynamical systems is to decide which pairs $([\sigma], M)$ can be realized by Koopman representations. The most spectacular, still unsolved, is the Banach problem concerning $([\lambda_{\mathbb{T}}], M \equiv 1)$. Another important problem is to give complete spectral classification in some classes of dynamical systems (classically, it was done in the theory of Kolmogorov and Gaussian dynamical systems). We will also see how spectral properties of dynamical systems can determine their statistical (ergodic) properties; a culmination given by the fact that a spectral isomorphism may imply measure-theoretic similitude (discrete spectrum case, Gaussian–Kronecker case). We conjecture that a dynamical system $\mathcal{T}$ either is spectrally determined or there are uncountably many pairwise non-isomorphic systems spectrally isomorphic to $\mathcal{T}$.

We could also consider Koopman representations in L^p for $1 \leq p \neq 2$. However whenever $W\colon L^p(X, \mathcal{B}, \mu) \to L^p(Y, C, \nu)$ is a surjective isometry and $W \circ U_{T_a} = U_{S_a} \circ W$ for each $a \in \mathbb{A}$ then by the Lamperti Theorem (e. g. [172]) the isometry W has to come from a nonsingular pointwise map $R\colon Y \to X$ and, by ergodicity, R "preserves" the measure ν and hence establishes a measure-theoretic isomorphism [94] (see also [127]). Thus spectral classification of such Koopman representations goes back to the measure-theoretic classification of dynamical systems, so it looks hardly interesting. This does not mean that there are no interesting dynamical questions for $p \neq 2$. Let us mention still open Thouvenot's question (formulated in the 1980s) for $\mathbb{Z}$-actions: *For every ergodic* T *acting on* $(X, \mathcal{B}, \preceq)$, *does there exist* $f \in L^1(X, \mathcal{B}, \mu)$ *such that the closed linear span of* $f \circ T^n$, $n \in \mathbb{Z}$ *equals* $L^1(X, \mathcal{B}, \mu)$?

Iwanik [79,80] proved that if T is a system with positive entropy then its L^p-multiplicity is ∞ for all $p > 1$. Moreover, Iwanik and de Sam Lazaro [85] proved that for Gaussian systems (they will be considered in Sect. "Spectral Theory of Dynamical Systems of Probabilistic Origin") the L^p-multiplicities are the same for all $p > 1$ (see also [137]).

Markov Operators, Joinings and Koopman Representations, Disjointness and Spectral Disjointness, Entropy

We would like to emphasize that spectral theory is closely related to the theory of joinings (see ► Joinings in Ergodic Theory for needed definitions). The elements ρ of the set $J(\mathcal{S},\mathcal{T})$ of joinings of two $\mathbb{A}$-actions $\mathcal{S}$ and $\mathcal{T}$ are in a 1-1 correspondence with Markov operators $J = J_\rho$ between the L^2-spaces equivariant with the corresponding Koopman representations (see the glossary and ► Joinings in Ergodic Theory). The set of ergodic self-joinings of an ergodic $\mathbb{A}$-action $\mathcal{T}$ is denoted by $J_2^e(\mathcal{T})$.

Each Koopman representation $\mathcal{U}_{\mathcal{T}}$ consists of Markov operators (indeed, U_{T_a} is clearly a Markov operator). In fact, if $U \in U(L^2(X,\mathcal{B},\mu))$ is Markov then it is of the form U_R, where $R \in Aut(X,\mathcal{B},\mu)$ [133]. This allows us to see Koopman representations as unitary Markov representations, but since a spectral isomorphism does not "preserve" the set of Markov operators, spectrally isomorphic systems can have drastically different sets of self-joinings.

We will touch here only some aspects of interactions (clearly, far from completeness) between the spectral theory and the theory of joinings.

In order to see however an example of such interactions let us recall that the simplicity of eigenvalues for ergodic systems yields a short "joining" proof of the classical isomorphism theorem of Halmos-von Neumann in the discrete spectrum case: *Assume that* $\mathcal{S} = (S_\dashv)_{\dashv\in\mathbb{A}}$ *and* $\mathcal{T} = (T_a)_{a\in\mathbb{A}}$ *are ergodic* $\mathbb{A}$*-actions on* $(X,\mathcal{B},\mu)$ *and* (Y,C,ν) *respectively. Assume that both Koopman representations have purely discrete spectrum and that their group of eigenvalues are the same. Then* $\mathcal{S}$ *and* $\mathcal{T}$ *are measure-theoretically isomorphic*. Indeed, each ergodic joining of $\mathcal{T}$ and $\mathcal{S}$ is the graph of an isomorphism of these two systems (see [127]; see also Goodson's proof in [66]). Another example of such interactions appear in the study Rokhlin's multiple mixing problem and its relation with the *pairwise independence property* (PID) for joinings of higher order. We will not deal with this subject here, referring the reader to ► Joinings in Ergodic Theory (see however Sect. "Lifting Mixing Properties").

Following [60], two $\mathbb{A}$-actions $\mathcal{S}$ and $\mathcal{T}$ are called *disjoint* provided the product measure is the only element in $J(\mathcal{S},\mathcal{T})$. It was already noticed in [72] that spectrally disjoint systems are disjoint in the Furstenberg sense; indeed, $Im(J_\rho|_{L_0^2}) = \{0\}$ since $\sigma_{\mathcal{T},J_\rho f} \ll \sigma_{\mathcal{S},f}$.

Notice that whenever we take $\rho \in J_2^e(\mathcal{T})$ we obtain a new ergodic $\mathbb{A}$-action $(T_a \times T_a)_{a\in\mathbb{A}}$ defined on the probability space $(X \times X, \rho)$. One can now ask how close spectrally to $\mathcal{T}$ is this new action? It turns out that except of the obvious fact that the marginal σ-algebras are factors, $(\mathcal{T} \times \mathcal{T}, \rho)$ can have other factors spectrally disjoint with $\mathcal{T}$: the most striking phenomenon here is a result of Smorodinsky and Thouvenot [198] (see also [29]) saying that each zero entropy system is a factor of an ergodic self-joining system of a fixed Bernoulli system (Bernoulli systems themselves have countable Haar spectrum). The situation changes if $\rho = \mu \otimes \mu$. In this case for $f, g \in L^2(X,\mathcal{B},\mu)$ the spectral measure of $f \otimes g$ is equal to $\sigma_{\mathcal{T},f} * \sigma_{\mathcal{T},g}$. A consequence of this observation is that an ergodic action $\mathcal{T} = (T_a)_{a\in\mathbb{A}}$ is weakly mixing (see the glossary) if and only if the product measure $\mu \otimes \mu$ is an ergodic self-joining of $\mathcal{T}$.

The entropy which is a basic measure-theoretic invariant does not appear when we deal with spectral properties. We will not give here any formal definition of entropy for amenable group actions referring the reader to [153]. Assume that $\mathbb{A}$ is countable and discrete. We always assume that $\mathbb{A}$ is Abelian, hence it is amenable. For each dynamical system $\mathcal{T} = (T_a)_{a\in\mathbb{A}}$ acting on $(X,\mathcal{B},\mu)$, we can find a largest invariant sub-σ field $\mathcal{P} \subset \mathcal{B}$, called the *Pinsker* σ-algebra, such that the entropy of the corresponding quotient system is zero. Generalizing the classical Rokhlin-Sinai Theorem (see also [97] for $\mathbb{Z}^d$-actions), Thouvenot (unpublished) and independently Dooley and Golodets [31] proved this theorem for groups even more general than those considered here: *The spectrum of* $\mathcal{U}_{\mathcal{T}}$ *on* $L^2(X,\mathcal{B},\mu) \ominus L^2(\mathcal{P})$ *is Haar with uniform infinite multiplicity*. This general result is quite intricate and based on methods introduced to entropy theory by Rudolph and Weiss [179] with a very surprising use of Dye's Theorem on orbital equivalence of all ergodic systems. For $\mathbb{A}$ which is not countable the same result was recently proved in [17] in case of unimodular amenable groups which are not increasing union of compact subgroups. It follows that spectral theory of dynamical systems essentially reduces to the zero entropy case.

Maximal Spectral Type of a Koopman Representation, Alexeyev's Theorem

Only few general properties of maximal spectral types of Koopman representations are known. The fact that a Koopman representation preserves the space of real functions implies that its maximal spectral type is the type of a symmetric (invariant under the map $\chi \mapsto \overline{\chi}$) measure.

Recall that the *Gelfand spectrum* $\sigma(\mathcal{U})$ of a representation $\mathcal{U} = (U_a)_{a\in\mathbb{A}}$ is defined as the of *approximative eigenvalues* of $\mathcal{U}$, that is $\sigma(\mathcal{U}) \ni \chi \in \widehat{\mathbb{A}}$ if for a sequence (x_n) bounded and bounded away from zero, $\|U_a x_n - \chi(a)x_n\| \to 0$ for each $a \in \mathbb{A}$. The spectrum is a closed subset in the topology of pointwise convergence, hence in the compact-open topology of $\widehat{\mathbb{A}}$. In case of $\mathbb{A} = \mathbb{Z}$, the above set $\sigma(U)$ is equal to $\{z \in \mathbb{C}\colon U - z \cdot Id$ is not invertible$\}$.

Assume now that $\mathbb{A}$ is countable and discrete (and Abelian). Then there exists a Fölner sequence $(B_n)_{n\geq 1}$ whose elements tile $\mathbb{A}$ [153]. Take a free and ergodic action $\mathcal{T} = (T_a)_{a\in\mathbb{A}}$ on $(X, \mathcal{B}, \mu)$. By [153] for each $\varepsilon > 0$ we can find a set $Y_n \in \mathcal{B}$ such that the sets $T_b Y_n$ are pairwise disjoint for $b \in B_n$ and $\mu(\bigcup_{b\in B_n} T_b Y_n) > 1 - \varepsilon$. For each $\chi \in \widehat{\mathbb{A}}$, by considering functions of the form $f_n = \sum_{b\in B_n} \chi(b) 1_{T_b Y_n}$ we obtain that $\chi \in \sigma(\mathcal{U}_{\mathcal{T}})$. It follows that the topological support of the maximal spectral type of the Koopman representation of a free and ergodic action is full [105,127,147]. The theory of Gaussian systems shows in particular that there are symmetric measures on the circle whose topological support is the whole circle but which cannot be maximal spectral types of Koopman representations.

An open well-known question remains whether an absolutely continuous measure ρ is the maximal spectral type of a Koopman representation if and only if ρ is equivalent to a Haar measure of $\widehat{\mathbb{A}}$ (this is unknown for $\mathbb{A} = \mathbb{Z}$).

Another interesting question was recently raised by A. Katok (private communication): *Is it true that the topological supports of all measures in a spectral sequence of a Koopman representation are full*? If the answer to this question is positive then for example the essential supremum of $M_{\mathcal{U}_{\mathcal{T}}}$ is the same on all balls of $\widehat{\mathbb{A}}$.

Notice that the fact that all spectral measures in a spectral sequence are symmetric means that $\mathcal{U}_{\mathcal{T}}$ is isomorphic to $\mathcal{U}_{\mathcal{T}^{-1}}$. A. del Junco [89] showed that generically for $\mathbb{Z}$-actions, T and its inverse are not measure-theoretically isomorphic (in fact he proved disjointness).

Let $\mathcal{T}$ be an $\mathbb{A}$-action on $(X, \mathcal{B}, \mu)$. One can ask wether a "good" function can realize the maximal spectral type of $\mathcal{U}_{\mathcal{T}}$. In particular can we find a function $f \in L^\infty(X, \mathcal{B}, \mu)$ that realizes the maximal spectral type? The answer is given in the following general theorem (see [139]).

Theorem 3 (Alexeyev's Theorem) *Assume that $\mathcal{U} = (U_a)_{a\in\mathbb{A}}$ is a unitary representation of $\mathbb{A}$ in a separable Hilbert space H. Assume that $F \subset H$ is a dense linear subspace. Assume moreover that with some F-norm $\left\vert\!\left\vert \cdot \right\vert\!\right\vert$ – stronger than the norm $\|\cdot\|$ given by the scalar product – F becomes a Fréchet space. Then, for each spectral measure σ ($\ll \sigma_{\mathcal{U}}$) there exists $y \in F$ such that $\sigma_y \gg \sigma$. In particular, there exists $y \in F$ that realizes the maximal spectral type.*

By taking $H = L^2(X, \mathcal{B}, \mu)$ and $F = L^\infty(X, \mathcal{B}, \mu)$ we obtain the positive answer to the original question. Alexeyev [14] proved the above theorem for $\mathbb{A} = \mathbb{Z}$ using analytic functions. Refining Alexeyev's original proof, Frączek [52] showed the existence of a sufficiently regular function realizing the maximal spectral type depending only on the "regularity" of the underlying probability space, e. g. when X is a compact metric space (compact manifold) then one can find a continuous (smooth) function realizing the maximal spectral type.

By the theory of systems of probabilistic origin (see Sect. "Spectral Theory of Dynamical Systems of Probabilistic Origin"), in case of simplicity of the spectrum, one can easily point out spectral measures whose types are not realized by (essentially) bounded functions. However, it is still an open question whether for each Koopman representation $\mathcal{U}_{\mathcal{T}}$ there exists a sequence $(f_i)_{i\geq 1} \subset L^\infty(X, \mathcal{B}, \mu)$ such that the sequence $(\sigma_{f_i})_{i\geq 1}$ is a spectral sequence for $\mathcal{U}_{\mathcal{T}}$. For $\mathbb{A} = \mathbb{Z}$ it is unknown whether the maximal spectral type of a Koopman representation can be realized by a characteristic function.

Spectral Theory of Weighted Operators

We now pass to the problem of possible essential values for the multiplicity function of a Koopman representation. However, one of known techniques is a use of cocycles, so before we tackle the multiplicity problem, we will go through recent results concerning spectral theory of compact group extensions automorphisms which in turn entail a study of weighted operators (see the glossary).

Assume that T is an ergodic automorphism of a standard Borel probability space $(X, \mathcal{B}, \mu)$. Let $\xi\colon X \to \mathbb{T}$ be a measurable function and let $V = V_{\xi,T}$ be the corresponding weighted operator. To see a connection of weighted operators with Koopman representations of compact group extensions consider a compact (metric) Abelian group G and a cocycle $\varphi\colon X \to G$. Then U_{T_φ} (see the glossary) acts on $L^2(X \times G, \mu \otimes \lambda_G)$. But

$$L^2(X\times G, \mu\otimes\lambda_G) = \bigoplus_{\chi\in\widehat{G}} L_\chi, \quad \text{where } L_\chi = L^2(X,\mu)\otimes\chi,$$

where L_χ is a U_{T_φ}-invariant (clearly, closed) subspace. Moreover, the map $f \otimes \chi \mapsto f$ settles a unitary isomorphism of $U_{T_\varphi}|_{L_\chi}$ with the operator $V_{\chi\circ\varphi,T}$. Therefore, spectral analysis of such Koopman representations reduces to the spectral analysis of weighted operators $V_{\chi\circ\varphi,T}$ for all $\chi \in \widehat{G}$.

Maximal Spectral Type of Weighted Operators over Rotations

The spectral analysis of weighted operators $V_{\xi,T}$ is especially well developed in case of rotations, i. e. when we additionally assume that T is an ergodic rotation on a compact monothetic group X: $Tx = x + x_0$, where x_0 is a topologically cyclic element of X (and μ will stand for Haar measure λ_x of X). In this case Helson's analysis [74] applies (see also [68,82,127,160]) leading to the following conclusions:

- The maximal spectral type $\sigma_{V_{\xi,T}}$ is either discrete or continuous.
- When $\sigma_{V_{\xi,T}}$ is continuous it is either singular or Lebesgue.
- The spectral multiplicity of $V_{\xi,T}$ is uniform.

We now pass to a description of some results in case when $Tx = x + \alpha$ is an irrational rotation on the additive circle $X = [0, 1)$. It was already noticed in the original paper by Anzai [16] that when $\xi\colon X \to \mathbb{T}$ is an affine cocycle ($\xi(x) = \exp(nx + c)$, $0 \neq n \in \mathbb{Z}$) then $V_{\xi,T}$ has a Lebesgue spectrum. It was then considered by several authors (originated by [123], see also [24,26]) to which extent this property is stable when we perturb our cocycle. Since the topological degree of affine cocyles is different from zero, when perturbing them we consider smooth perturbations by cocycles of degree zero.

Theorem 4 ([82]) *Assume that $Tx = x + \alpha$ is an irrational rotation. If $\xi\colon [0, 1) \to \mathbb{T}$ is of non-zero degree, absolutely continuous, with the derivative of bounded variation then $V_{\xi,T}$ has a Lebesgue spectrum.*

In the same paper, it is noticed that if we drop the assumption on the derivative then the maximal spectral type of $V_{\xi,T}$ is a Rajchman measure (i. e. its Fourier transform vanishes at infinity). It is still an open question, whether one can find ξ absolutely continuous with non-zero degree and such that $V_{\xi,T}$ has singular spectrum. "Below" absolute continuity, topological properties of the cocycle seem to stop playing any role – in [82] a continuous, degree 1 cocycle ξ of bounded variation is constructed such that $\xi(x) = \eta(x)/\eta(Tx)$ for a measurable $\eta\colon [0, 1) \to \mathbb{T}$ (that is ξ is a *coboundary*) and therefore $V_{\xi,T}$ has purely discrete spectrum (it is isomorphic to U_T). For other results about Lebesgue spectrum for Anzai skew products see also [24,53,81] (in [53] $\mathbb{Z}^d$-actions of rotations and so called winding numbers instead of topological degree are considered).

When the cocycle is still smooth but its degree is zero the situation drastically changes. Given an absolutely continuous function $f\colon [0, 1) \to \mathbb{R}$ M. Herman [76], using the Denjoy–Koksma inequality (see e. g. [122]), showed that $f_0^{(q_n)} \to 0$ uniformly (here $f_0 = f - \int_0^1 f \, d\lambda_{[0,1)}$ and (q_n) stands for the sequence of denominators of α). It follows that $T_{e^{2\pi i f}}$ is rigid and hence has a singular spectrum. B. Fayad [37] shows that this result is no longer true if one dimensional rotation is replaced by a multi-dimensional rotation (his counterexample is in the analytic class). See also [130] for the singularity of spectrum for functions f whose Fourier transform satisfies $o(1/|n|)$ condition or to [84], where it is shown that sufficiently small variation implies singularity of the spectrum.

A natural class of weighted operators arises when we consider Koopman operators of rotations on nil-manifolds. We only look at the particular example of such a rotation on a quotient of the Heisenberg group $(\mathbb{R}^3, *)$ (a general spectral theory of nil-actions was mainly developed by W. Parry [157]) – these actions have countable Lebesgue spectrum in the orthocomplement of the subspace of eigenfunctions) that is take the nil-manifold $\mathbb{R}^3/_*\mathbb{Z}^3$ on which we define the nil-rotation $S((x, y, z) * \mathbb{Z}^3) = (\alpha, \beta, 0)*(x, y, z)*\mathbb{Z}^3 = \big(x + \alpha, y + \beta, z + \alpha y\big)* \mathbb{Z}^3$, where α, β and 1 are rationally independent. It can be shown that S is isomorphic to the skew product defined on $[0, 1)^2 \times \mathbb{T}$ by

$$T_\varphi\colon (x, y, z) \mapsto \Big(x + \alpha, y + \beta, z \cdot e^{2\pi i\varphi(x,y)}\Big) ,$$

where $\varphi(x, y) = \alpha y - \psi(x + \alpha, y + \beta) + \psi(x, y)$ with $\psi(x, y) = x[y]$. Since nil-cocycles can be considered as a certain analog of affine cocycles for one-dimensional rotations, it would be nice to explain to what kind of perturbations the Lebesgue spectrum property is stable.

Yet another interesting problem which is related to the spectral theory of extensions given by so called *Rokhlin cocycles* (see Sect. "Rokhlin Cocycles") arises, when given $f\colon [0, 1) \to \mathbb{R}$, we want to describe spectrally the one-parameter set of weighted operators $W_c := V_{e^{2\pi i c f},T}$; here T is a fixed irrational rotation by α. As proved by quite sophisticated arguments in [84], if we take $f(x) = x$ then for all non-integer $c \in \mathbb{R}$ the spectrum of W_c is continuous and singular (see also [68] and [145] where some special α's are considered). It has been open for some time if at all one can find $f\colon [0, 1) \to \mathbb{R}$ such that for each $c \neq 0$, the operator W_c has a Lebesgue spectrum. The positive answer is given in [205]: for example if $f(x) = x^{-(2+\varepsilon)}$ ($\varepsilon > 0$) and α has bounded partial quotients then W_c has a Lebesgue spectrum for all $c \neq 0$. All functions with such a property considered in [205] are non-integrable. It would be interesting to find an integrable f with the above property.

We refer to [66] and the references therein for further results especially for transformations of the form $(x, y) \mapsto (x + \alpha, 1_{[0,\beta)}(x) + y)$ on $[0, 1) \times \mathbb{Z}/2\mathbb{Z}$. Recall however that earlier Katok and Stepin [104] used this kind of transformations to give a first counterexample to the Kolmogorov group property (see the glossary) for the spectrum.

The Multiplicity Problem for Weighted Operators over Rotations

In case of perturbations of affine cocycles, this problem was already raised by Kushnirenko [123]. Some significant results were obtained by M. Guenais. Before we state her results let us recall a useful criterion to find an upper bound for the multiplicity: *If there exist $c > 0$ and a sequence $(F_n)_{n \geq 1}$ of cyclic subspaces of H such that for each $y \in H$, $\|y\| = 1$ we have $\liminf_{n\to\infty} \|proj_{F_n} y\|^2 \geq c$, then $esssup(M_U) \leq 1/c$* which follows from a well-known lemma of Chacon [23,26,111,127]. Using this and a technique which is close to the idea of local rank one (see [44,111]) M. Guenais [69] proved a series of results on multiplicity which we now list.

Theorem 5 *Assume that $Tx = x + \alpha$ and let $\xi \colon [0, 1) \to \mathbb{T}$ be a cocycle.*

(i) If $\xi(x) = e^{2\pi icx}$ then $M_{V_{\xi,T}}$ is bounded by $|c| + 1$.
(ii) If ξ is absolutely continuous and ξ is of topological degree zero, then $V_{\xi,T}$ has a simple spectrum.
(iii) if ξ is of bounded variation, then $M_{V_{\xi,T}} \leq \max(2, 2\pi\, Var(\xi)/3)$.

Remarks on the Banach Problem

We already mentioned in Introduction the Banach problem in ergodic theory, which is simply the question whether there exists a Koopman representation for $\mathbb{A} = \mathbb{Z}$ with simple Lebesgue spectrum. In fact no example of a Koopman representation with Lebesgue spectrum of finite multiplicity is known. Helson and Parry [75] asked for the validity of a still weaker version: *Can one construct T such that U_T has a Lebesgue component in its spectrum whose multiplicity is finite*? Quite surprisingly in [75] they give a general construction yielding for each ergodic T a cocycle $\varphi \colon X \to \mathbb{Z}/2\mathbb{Z}$ such that the unitary operator U_{T_φ} has a Lebesgue spectrum in the orthocomplement of functions depending only on the X-coordinate. Then Mathew and Nadkarni [144] gave examples of cocycles over so called dyadic adding machine for which the multiplicity of the Lebesgue component was equal to 2. In [126] this was generalized to so called *Toeplitz $\mathbb{Z}/2\mathbb{Z}$-extensions* of adding machines: for each even number k we can find a two-point extension of an adding machine so that the multiplicity of the Lebesgue component is k. Moreover, it was shown that Mathew and Nadkarni's constructions from [144] in fact are close to systems arising from number theory (like the famous Rudin–Shapiro sequence, e. g. [160]), relating the result about multiplicity of the Lebesgue component to results by Kamae [96] and Queffelec [160]. Independently of [126], Ageev [8] showed that one can construct 2-point extensions of the Chacon transformation realizing (by taking powers of the extension) each even number as the multiplicity of the Lebesgue component. Contrary to all previous examples, Ageev's constructions are weakly mixing.

Still an open question remains whether for $\mathbb{A} = \mathbb{Z}$ one can find a Koopman representation with the Lebesgue component of multiplicity 1 (or even odd).

In [70], M. Guenais studies the problem of Lebesgue spectrum in the classical case of Morse sequences (see [107] as well as [124], where the problem of spectral classification in this class is studied). All dynamical sytems arising from Morse sequences have simple spectra [124]. It follows that if a Lebesgue component appears in a Morse dynamical system, it has multiplicity one. Guenais [70] using a Riesz product technique relates the Lebesgue spectrum problem with the still open problem of whether a construction of "flat" trigonometric polynomials with coefficients ±1 is possible. However, it is proved in [70] that such a construction can be carried out on some compact Abelian groups and it leads, for an Abelian countable torsion group $\mathbb{A}$, to a construction of an ergodic action of $\mathbb{A}$ with simple spectrum and a Haar component in its spectrum.

Lifting Mixing Properties

We now give one more example of interactions between spectral theory and joinings (see Introduction) that gives rise to a quick proof of the fact that r-fold mixing property of T ($r \geq 2$) lifts to a weakly mixing compact group extension T_φ (the original proof of this fact is due to D. Rudolph [175]). Indeed, to prove r-fold mixing for a mixing(=2-mixing) transformation S (acting on (Y, C, ν)) one has to prove that each weak limit of off-diagonal self-joinings (given by powers of S, see ▶ Joinings in Ergodic Theory) of order r is simply the product measure $\nu^{\otimes r}$. We need also a Furstenberg's lemma [62] about relative unique ergodicity (RUE) of compact group extensions T_φ: *If $\mu \otimes \lambda_G$ is an ergodic measure for T_φ then it is the only (ergodic) invariant measure for T_φ whose projection on the first coordinate is μ.* Now the result about lifting r-fold mixing to compact group extensions follows directly from the fact that whenever T_φ is weakly mixing, $(\mu \otimes \lambda_G)^{\otimes r}$ is an ergodic measure (this approach was shown to me by A. del Junco). In particular if T is mixing and T_φ is weakly mixing then for each $\chi \in \widehat{G} \setminus \{1\}$, the maximal spectral type of $V_{\chi\circ\varphi,T}$ is Rajchman.

See Sect. "Rokhlin Cocycles" for a generalization of the lifting result to Rokhlin cocycle extensions.

The Multiplicity Function

In this chapter only $\mathbb{A} = \mathbb{Z}$ is considered (for other groups, even for $\mathbb{R}$, much less is known; see however the case of so called *product* $\mathbb{Z}^d$*-actions* [50]). Contrary to the case of maximal spectral type, it is rather commonly believed that there are no restrictions for the set of essential values of Koopman representations.

Cocycle Approach

We will only concentrate on some results of the last twenty years. In 1983, E.A. Robinson [164] proved that for each $n \geq 1$ there exists an ergodic transformation whose maximal spectral multiplicity is n. Another important result was proved in [165] (see also [98]), where it is shown that given a finite set $M \subset \mathbb{N}$ containing 1 and closed under the least common multiple one can find (even a weakly mixing) T so that the set of essential values of the multiplicity function equals M. This result was then extended in [67] to infinite sets and finally in [125] (see also [11]) to all subsets $M \subset \mathbb{N}$ containing 1. In fact, as we have already noticed in the previous section the spectral theory for compact Abelian group extensions is reduced to a study of weighted operators and then to comparing maximal spectral types for such operators. This leads to sets of the form

$$M(G, v, H) = \Big\{\sharp(\{\chi \circ v^i : i \in \mathbb{Z}\} \cap anih(H)) : \chi \in anih(H)\Big\}$$

($H \subset G$ is a closed subgroup and v is a continuous group automorphism of G). Then an algebraic lemma has been proved in [125] saying that each set M containing 1 is of the form $M(G, v, H)$ and the techniques to construct "good" cocycles and a passage to "natural factors" yielded the following: *For each $M \subset \{1, 2, \ldots\} \subset \{\infty\}$ containing 1 there exists an ergodic automorphim such that the set of essential values for its Koopman representation equals M.* See also [166] for the case of non-Abelian group extensions.

A similar in spirit approach (that means, a passage to a family of factors) is present in a recent paper of Ageev [13] in which he first applies Katok's analysis (see [98,102]) for spectral multiplicities of the Koopman representation associated with Cartesian products $T^{\times k}$ for a generic transformation T. In a natural way this approach leads to study multiplicities of tensor products of unitary operators. Roughly, the multiplicity is computed as the number of atoms (counted modulo obvious symmetries) for conditional measures (see [98]) of a product measure over its convolution. Ageev [13] proved that for a typical automorphism T the set of the values of the multiplicity function for $U_{T^{\times k}}$ equals $\{k, k(k-1), \ldots, k!\}$ and then he just passes to "natural" factors for the Cartesian products by taking sets invariant under a fixed subgroup of permutations of coordinates. In particular, he obtains all sets of the form $\{2, 3, \ldots, n\}$ on L^2_0. He also shows that such sets of multiplicities are realizable in the category of mixing transformations.

Rokhlin's Uniform Multiplicity Problem

The Rokhlin multiplicity problem (see the recent book by Anosov [15]) was, given $n \geq 2$, to construct an ergodic transformation with uniform multiplicity n on L^2_0. A solution for $n = 2$ was independently given by Ageev [9] and Ryzhikov [188] (see also [15] and [66]) and in fact it consists in showing that for some T (actually, any T with simple spectrum for which $1/2(Id + U_T)$ is in the weak operator closure of the powers of U_T will do) the multiplicity of $T \times T$ is uniformly equal to 2 (see also Sect. "Future Directions").

In [12], Ageev proposed a new approach which consists in considering actions of "slightly non-Abelian" groups; and showing that for a "typical" action of such a group a fixed "direction" automorphism has a uniform multiplicity. Shortly after publication of [12], Danilenko [27], following Ageev's approach, considerably simplified the original proof. We will present Danilenko's arguments.

Fix $n \geq 1$. Denote $\overline{e}_i = (0, \ldots, 1, \ldots, 0) \in \mathbb{Z}^n$, $i = 1, \ldots, n$. We define an automorphism L of $\mathbb{Z}^n$ setting $L(\overline{e}_i) = \overline{e}_{i+1}$, $i = 1, \ldots, n-1$ and $L(\overline{e}_n) = \overline{e}_1$. Using L we define a semi-direct product $G\colon = \mathbb{Z}^n \rtimes \mathbb{Z}$ defining multiplication as $(u, k) \cdot (w, l) = (u + L^k w, k + l)$. Put $e_0 = (0, 1)$, $e_i = (\overline{e}_i, 0)$, $i = 1, \ldots, n$ (and $Le_i = (L\overline{e}_i, 0)$). Moreover, denote $e_{n+1} = e_0^n = (0, n)$. Notice that $e_0 \cdot e_i \cdot e_0^{-1} = Le_i$ for $i = 1, \ldots, n$ $(L(e_{n+1}) = e_{n+1})$.

Theorem 6 (Ageev, Danilenko) *For every unitary representation $\mathcal{U}$ of G in a separable Hilbert space H, for which $U_{e_1 - L^r e_1}$ has no non-trivial fixed points for $1 \leq r < n$, the essential values of the multiplicity function for $U_{e_{n+1}}$ are contained in the set of multiples of n. If, in addition, U_{e_0} has a simple spectrum, then $U_{e_{n+1}}$ has uniform multiplicity n.*

It is then a certain work to show that the assumption of the second part of the theorem is satisfied for a typical action of the group G. Using a special (C, F)-construction with all the cut-and-stack parameters explicit Danilenko [27] was also able to show that each set of the form $k \cdot M$, where $k \geq 1$ and M is an arbitrary subset of natural numbers containing 1, is realizable as the set of essential values of a Koopman representation.

Some other constructions based on the solution of the Rokhlin problem for $n = 2$ and the method of [125] are presented in [103] leading to sets different than those pointed above; these sets contain 2 as their minimum.

Rokhlin Cocycles

We consider now a certain class of extensions which should be viewed as a generalization of the concept of compact group extensions. We will focus on $\mathbb{Z}$-actions only.

Assume that T is an ergodic automorphism of $(X, \mathcal{B}, \mu)$. Let G be a l.c.s.c. Abelian group. Assume that this group acts on (Y, C, ν), that is we have a G-action $S = (S_g)_{g \in G}$ on (Y, C, ν). Let $\varphi\colon X \to G$ be a cocycle. We then define an automorphism $T_{\varphi,S}$ of the space $(X \times Y, \mathcal{B} \otimes C, \mu \otimes \nu)$ by

$$T_{\varphi,S}(x, y) = (Tx, S_{\varphi(x)}(y)).$$

Such an extension is called a *Rokhlin cocycle extension* (the map $x \mapsto S_{\varphi(x)}$ is called a *Rokhlin cocycle*). Such an operation generalizes the case of compact group extensions; indeed, when G is compact the action of G on itself by rotations preserves Haar measure. (It is quite surprising, that when only we admit G non-Abelian, then, as shown in [28], **each** ergodic extension of T has a form of a Rokhlin cocycle extension.) Ergodic and spectral properties of such extensions are examined in several papers: [63,65,129,131,132,133,167,176]. Since in these papers rather joining aspects are studied (among other things in [129] Furstenberg's RUE lemma is generalized to this new context), we will mention here only few results, mainly spectral, following [129] and [133]. We will constantly assume that G is non-compact. As $\varphi\colon X \to G$ is then a cocycle with values in a non-compact group, the theory of such cocycles is much more complicated (see e. g. [193]), and in fact the theory of Rokhlin cocycle extensions leads to interesting interactions between classical ergodic theory, the theory of cocycles and the theory of non-singular actions arising from cocycles taking values in non-compact groups – especially, the Mackey action associated to φ plays a crucial role here (see the problem of invariant measures for $T_{\varphi,S}$ in [132] and [28]); see also monographs [1,98,101,193]. Especially, two Borel subgroups of $\widehat{G}$ are important here:

$$\Sigma_\varphi = \{\chi \in \widehat{G}\colon \chi \circ \varphi = c \cdot \xi/\xi \circ T \text{ for a measurable } \xi\colon X \to \mathbb{T} \text{ and } c \in \mathbb{T}\}\,.$$

and its subgroup Λ_φ given by $c = 1$. Λ_φ turns out to be the group of L^∞-eigenvalues of the Mackey action (of G) associated to the cocycle φ. This action is the quotient action of the natural action of G (by translations on the second coordinate) on the space of ergodic components of the skew product T_φ – the Mackey action is (in general) not measure-preserving, it is however non-singular. We refer the reader to [2,78,147] for other properties of those subgroups.

Theorem 7 ([132,133])

(i) $\sigma_{T_{\varphi,S}}|_{L^2(X\times Y,\mu\otimes\nu)\ominus L^2(X,\mu)} = \int_{\widehat{G}} \sigma_{V_{\chi\circ\varphi,T}}\, d\sigma_S$.

(ii) *$T_{\varphi,S}$ is ergodic if and only if T is ergodic and $\sigma_S(\Lambda_\varphi) = 0$.*

(iii) *$T_{\varphi,S}$ is weakly mixing if and only if T is weakly mixing and S has no eigenvalues in Σ_φ.*

(iv) *if T is mixing, S is mildly mixing, φ is recurrent and not cohomologous to a cocycle with values in a compact subgroup of G then $T_{\varphi,S}$ remains mixing.*

(v) *If T is r-fold mixing, φ is recurrent and $T_{\varphi,S}$ is mildly mixing then $T_{\varphi,S}$ is also r-fold mixing.*

(vi) *If T and R are disjoint, the cocycle φ is ergodic and S is mildly mixing then $T_{\varphi,S}$ remains disjoint with R.*

Let us recall [61,195] that an $\mathbb{A}$-action $S = (S_a)_{a\in\mathbb{A}}$ is mildly mixing (see the glossary) if and only if the $\mathbb{A}$-action $(S_a \times \tau_a)_{a\in\mathbb{A}}$ remains ergodic for every properly ergodic non-singular $\mathbb{A}$-action $\tau = (\tau_a)_{a\in\mathbb{A}}$.

Coming back to Smorodinsky–Thouvenot's result about factors of ergodic self-joinings of a Bernoulli automorphism we would like to emphasize here that the disjointness result (*vi*) above was used in [132] to give an example of an automorphism which is disjoint from all weakly mixing transformations but which has an ergodic self-joining whose associated automorphism has a non-trivial weakly mixing factor. In a sense this is opposed to Smorodinsky–Thouvenot's result as here from self-joinings we produced a "more complicated" system (namely the weakly mixing factor) than the original system.

It would be interesting to develop the theory of spectral multiplicity for Rokhlin cocycle extensions as it was done in the case of compact group extensions. However a difficulty is that in the compact group extension case we deal with a countable direct sum of representations of the form

$V_{\chi\circ\varphi,T}$ while in the non-compact case we have to consider an integral of such representations.

Rank-1 and Related Systems

Although properties like mixing, weak (and mild) mixing as well as ergodicity, are clearly spectral properties, they have "good" measure-theoretic formulations (expressed by a certain behavior on sets). Simple spectrum property is another example of a spectral property, and it was a popular question in the 1980s whether simple spectrum property of a Koopman representation can be expressed in a more "measure-theoretic" way. We now recall rank-1 concept which can be seen as a notion close to Katok's and Stepin's theory of cyclic approximation [104] (see also [26]).

Assume that T is an automorphism of a standard probability Borel space $(X, \mathcal{B}, \mu)$. T is said to have *rank one* property if there exists an increasing sequence of Rokhlin towers tending to the partition into points (a *Rokhlin tower* is a family $\{F, TF, \dots, T^{n-1}F\}$ of pairwise disjoint sets, while "tending to the partition into points" means that we can approximate every set in $\mathcal{B}$ by unions of levels of towers in the sequence). Baxter [20] showed that the maximal spectral type of such a T is realized by a characteristic function. Since the cyclic space generated by the characteristic function of the base contains characteristic functions of all levels of the tower, by the definition of rank one, the increasing sequence of cyclic spaces tends to the whole L^2-space, therefore rank one property implies simplicity of the spectrum for the Koopman representation. It was a question for some time whether rank-1 is just a characterization of simplicity of the spectrum, disproved by del Junco [88]. We refer the reader to [46] as a good source for basic properties of rank-1 transformations.

Similarly to the rank one property, one can define *finite rank* automorphisms (simply by requiring that an approximation is given by a sequence of a fixed number of towers) – see e. g. [152], or even, a more general property, namely the *local rank one* property can be defined, just by requiring that the approximating sequence of single towers fills up a fixed fraction of the space (see [44,111]). Local rank one (so the more finite rank) property implies finite multiplicity [111]. Mentzen [146] showed that for each $n \geq 1$ one can construct an automorphism with simple spectrum and having rank n; in [138] there is an example of a simple spectrum automorphism which is not of local rank one. Ferenczi [45] introduced the notion of funny rank one (approximating towers are Rokhlin towers with "holes"). Funny rank one also implies simplicity of the spectrum. An example is given in [45] which is even not loosely Bernoulli (see Sect. "Inducing and Spectral Theory", we recall that local rank one property implies loose Bernoullicity [44]).

The notion of AT (see the glossary) has been introduced by Connes and Woods [25]. They proved that AT property implies zero entropy. They also proved that funny rank one automorphisms are AT. In [32] it is proved that the system induced by the classical Morse-Thue system is AT (it is an open question by S. Ferenczi whether this system has funny rank one property). A question by Dooley and Quas is whether AT implies funny rank one property. AT property implies "simplicity of the spectrum in L^1" which we already considered in Introduction (a "generic" proof of this fact is due to J.-P. Thouvenot).

A persistent question was formulated in the 1980s whether rank one itself is a spectral property. In [49] the authors maintained that this is not the case, based on an unpublished preprint of the first named author of [49] in which there was a construction of a Gaussian–Kronecker automorphism (see Sect. "Spectral Theory of Dynamical Systems of Probabilistic Origin") having rank-1 property. This latter construction turned out to be false. In fact de la Rue [181] proved that no Gaussian automorphism can be of local rank one. Therefore the question whether: *Rank one is a spectral property* remains one of the interesting open questions in that theory. Downarowicz and Kwiatkowski [33] proved that rank-1 is a spectral property in the class of systems generated by generalized Morse sequences.

One of the most beautiful theorems about rank-1 automorphisms is the following result of J. King [110] (for a different proof see [186]).

Theorem 8 (WCT) *If T is of rank one then for each element S of the centralizer $C(T)$ of T there exists a sequence (n_k) such that $U_T^{n_k} \to U_S$ strongly.*

A conjecture of J. King is that in fact for rank-1 automorphisms each indecomposable Markov operator $J = J_\rho$ ($\rho \in J_2^e(T)$) is a weak limit of powers of U_T (see [112], also [186]). To which extent the WCT remains true for actions of other groups is not clear. In [214] the WCT is proved in case of rank one flows, however the main argument seems to be based on the fact that a rank one flow has a non-zero time automorphism T_{t_0} which is of rank one, which is not true. After the proof of the WCT by Ryzhikov in [186] there is a remark that the rank one flow version of the theorem can be proved by a word for word repetition of the arguments. He also proves that if the flow $(T_t)_{t\in\mathbb{R}}$ is mixing, then T_1 does not have finite rank. On the other hand, for $\mathbb{A} = \mathbb{Z}^2$, Downarowicz and Kwiatkowski [34] gave recently a counterexample to the WCT.

Even though it looks as if rank one construction is not complicated, mixing in this class is possible; historically the first mixing constructions were given by D. Ornstein [151] in 1970, using probability type arguments for a choice of spacers. Once mixing was shown, the question arose whether absolutely continuous spectrum is also possible, as this would give automatically the positive answer to the Banach problem. However Bourgain [21], relating spectral measures of rank one automorphisms with some classical constructions of Riesz product measures, proved that a certain subclass of Ornstein's class consists of automorphisms with singular spectrum (see also [5] and [6]). Since in Ornstein's class spacers are chosen in a certain "non-constructive" way, quite a lot of attention was devoted to the rank one automorphism defined by cutting a tower at the nth step into $r_n = n$ subcolumns of equal "width" and placing i spacers over the ith subcolumn. The mixing property conjectured by M. Smorodinsky, was proved by Adams [7] (in fact Adams proved a general result on mixing of a class of staircase transformations). Spectral properties of rank-1 transformations are also studied in [114], where the authors proved that whenever $\sum_{n=1}^{\infty} r_n^{-2} = +\infty$ (r_n stands for the number of subcolumns at the nth step of the construction of a rank-1 automorphism) then the spectrum is automatically singular. H. Abdalaoui [5] gives a criterion for singularity of the spectrum of a rank one transformation; his proof uses a central limit theorem. It seems that still the question whether rank one implies singularity of the spectrum remains the most important question of this theory.

We have already seen in Sect. "Spectral Theory of Weighted Operators" that for a special class of rank one systems, namely those with discrete spectra ([87]), we have a nice theory for weighted operators. It would be extremely interesting to find a rank one automorphism with continuous spectrum for which a substitute of Helson's analysis exists.

B. Fayad [39] constructs a rank one differentiable flow, as a special flow over a two-dimensional rotation. In [40] he gives new constructions of smooth flows with singular spectra which are mixing (with a new criterion for a Rajchman measure to be singular). In [35] a certain smooth change of time for an irrational flows on the 3-torus is given, so that the corresponding flow is partially mixing and has the local rank one property.

Spectral Theory of Dynamical Systems of Probabilistic Origin

Let us just recall that when $(Y_n)_{n=-\infty}^{\infty}$ is a stationary process then its distribution μ on $\mathbb{R}^{\mathbb{Z}}$ is invariant under the shift S on $\mathbb{R}^{\mathbb{Z}}$: $S((x_n)_{n\in\mathbb{Z}}) = (y_n)_{n\in\mathbb{Z}}$, where $y_n = x_{n+1}$, $n \in \mathbb{Z}$. In this way we obtain an automorphism S defined on $(\mathbb{R}^{\mathbb{Z}}, \mathcal{B}(\mathbb{R}^{\mathbb{Z}}), \mu)$. For each automorphism T we can find $f\colon X \to \mathbb{R}$ measurable such that the smallest σ-algebra making the stationary process $(f \circ T^n)_{n\in\mathbb{Z}}$ measurable is equal to $\mathcal{B}$, therefore, for the purpose of this article, by a system of probabilistic origin we will mean (S, μ) obtained from a stationary infinitely divisible process (see e. g. [142,192]). In particular, the theory of Gaussian dynamical systems is indeed a classical part of ergodic theory (e. g. [149,150,211,212]). If $(X_n)_{n\in\mathbb{Z}}$ is a stationary real centered Gaussian process and σ denotes the *spectral measure of the process*, i. e. $\widehat{\sigma}(n) = E(X_n \cdot X_0)$, $n \in \mathbb{Z}$, then by $S = S_\sigma$ we denote the corresponding Gaussian system on the shift space (recall also that for each symmetric measure σ on $\mathbb{T}$ there is exactly one stationary real centered Gaussian process whose spectral measure is σ). Notice that if σ has an atom, then in the cyclic space generated by X_0 there exists an eigenfunction Y for S_σ – if now S_σ were ergodic, $|Y|$ would be a constant function which is not possible by the nature of elements in $\mathbb{Z}(X_0)$. In what follows we assume that σ is continuous.

It follows that U_{S_σ} restricted to $\mathbb{Z}(X_0)$ is spectrally the same as $V = V^\sigma$ acting on $L^2(\mathbb{T}, \sigma)$, and we obtain that $(U_{S_\sigma}, L^2(\mathbb{R}^{\mathbb{Z}}, \mu_\sigma))$ can be represented as the symmetric Fock space built over $H = L^2(\mathbb{T}, \sigma)$ and $U_{S_\sigma} = F(V)$ – see the glossary ($H^{\odot n}$ is called the *n-th chaos*). In other words the spectral theory of Gaussian dynamical systems is reduced to the spectral theory of special tensor products unitary operators. Classical results (see [26]) which can be obtained from this point of view are for example the following:

(A) ergodicity implies weak mixing,
(B) the multiplicity function is either 1 or is unbounded,
(C) the maximal spectral type of U_{S_σ} is equal to $\exp(\sigma)$, hence Gaussian systems enjoy the Kolmogorov group property.

However we can also look at a Gaussian system in a different way, simply by noticing that the variables $e^{2\pi i f}$ (f is a real variable), where $f \in \mathbb{Z}(X_0)$ generate $L^2(\mathbb{R}^{\mathbb{Z}}, \mu_\sigma)$. Now calculating the spectral measure of $e^{2\pi i f}$ is not difficult and we obtain easily (C). Moreover, integrals of type $\int e^{2\pi i f_0} e^{2\pi i f_1 \circ T^n} e^{2\pi i f_2 \circ T^{n+m}}\, d\mu_\sigma$ can also be calculated, whence in particular we easily obtain Leonov's theorem on the multiple mixing property of Gaussian systems [141].

One of the most beautiful parts of the theory of Gaussian systems concerns ergodic properties of S_σ when σ is concentrated on a thin Borel set. Recall that a closed subset $K \subset \mathbb{T}$ is said to be a *Kronecker set* if each $f \in C(K)$ is a uniform limit of characters (restricted to K). Each Kronecker set has no rational relations. Gaussian–Kronecker automorphisms are, by definition, those Gaussian systems for which the measure σ (always assumed to be continuous) is concentrated on $K \cup \overline{K}$, K a Kronecker set. The following theorem has been proved in [51] (see also [26]).

Theorem 9 (Foiaş–Stratila Theorem) *If T is an ergodic automorphism and f is a real-valued element of L_0^2 such that the spectral measure σ_f is concentrated on $K \cup \overline{K}$, where K is a Kronecker set, then the process $(f \circ T^n)_{n\in\mathbb{Z}}$ is Gaussian.*

This theorem is indeed striking as it gives examples of weakly mixing automorphisms which are spectrally determined (like rotations). A relative version of the Foiaş–Stratila Theorem has been proved in [129].

The Foiaş–Stratila Theorem implies that whenever a spectral measure σ is Kronecker, it has no realization of the form σ_f with f bounded. We will see however in Sect. "Future Directions" that for some automorphisms T (having the SCS property) the maximal spectral type σ_T has the property that S_{σ_T} has a simple spectrum.

Gaussian–Kronecker automorphisms are examples of automorphisms with simple spectra. In fact, whenever σ is concentrated on a set without rational relations, then S_σ has a simple spectrum (see [26]). Examples of mixing automorphisms with simple spectra are known [149], however it is still unknown (Thouvenot's question) whether the Foiaş–Stratila property may hold in the mixing class. F. Parreau [154] using independent Helson sets gave an example of mildly mixing Gaussian system with the Foiaş–Stratila property.

In [165] there is a remark that the set of finite essential values of the multiplicity function of U_{S_σ} forms a (multiplicative) subsemigroup of $\mathbb{N}$. However, it seems that there is no "written" proof of this fact.

Joining theory of a class of Gaussian system, called GAG, is developed in [136]. A Gaussian automorphism S_σ with the Gaussian space $H \subset L_0^2(\mathbb{R}^{\mathbb{Z}}, \mu_\sigma)$ is called a GAG if for each ergodic self-joining $\rho \in J_2^e(S_\sigma)$ and arbitrary $f, g \in H$ the variable

$$(\mathbb{R}^{\mathbb{Z}} \times \mathbb{R}^{\mathbb{Z}}, \rho) \ni (x, y) \mapsto f(x) + g(y)$$

is Gaussian. For GAG systems one can describe the centralizer and factors, they turn out to be objects close to the probability structure of the system. One of the crucial observations in [136] was that all Gaussian systems with simple spectrum are GAG.

It is conjectured (J.P. Thouvenot) that in the class of zero entropy Gaussian systems the PID property holds true.

For the spectral theory of classical factors of a Gaussian system see [137]; also spectrally they share basic spectral properties of Gaussian systems. Recall that historically one of the classical factors namely the σ-algebra of sets invariant for the map

$$(\ldots, x_{-1}, x_0, x_1, \ldots) \mapsto (\ldots, -x_{-1}, -x_0, -x_1, \ldots)$$

was the first example with zero entropy and countable Lebesgue spectrum (indeed, we need a singular measure σ such that $\sigma * \sigma$ is equivalent to Lebesgue measure [150]). For factors obtained as functions of a stationary process see [83].

T. de la Rue [181] proved that Gaussian systems are never of local rank-1, however his argument does not apply to classical factors. We conjecture that Gaussian systems are disjoint from rank-1 automorphisms (or even from local rank-1 systems).

We now turn the attention to Poissonian systems (see [26] for more details). Assume that $(X, \mathcal{B}, \mu)$ is a standard Borel space, where μ is infinite, σ-finite. The new configuration space $\widetilde{X}$ is taken as the set of all countable subsets $\{x_i\colon i \geq 1\}$ of X. Once a set $A \in \mathcal{B}$, of finite measure is given one can define a map $N_A\colon \widetilde{X} \to \mathbb{N}(\cup\{\infty\})$ just counting the number of elements belonging to A. The measure-theoretic structure $(\widetilde{X}, \widetilde{\mathcal{B}}, \widetilde{\mu})$ is given so that the maps N_A become random variables with Poisson distribution of parameter $\mu(A)$ and such that whenever $A_1, \ldots, A_k \subset X$ are of finite measure and are pairwise disjoint then the variables $N_{A_1}, \ldots, N_{A_k}$ are independent.

Assume now that T is an automorphism of $(X, \mathcal{B}, \mu)$. It induces a natural automorphism on the space $(\widetilde{X}, \widetilde{\mathcal{B}}, \widetilde{\mu})$ defined by $\widetilde{T}(\{x_i\colon i \geq 1\} = \{Tx_i\colon i \geq 1\}$. The automorphism $\widetilde{T}$ is called the *Poisson suspension* of T (see [26]). Such a suspension is ergodic if and only if no set of positive and finite μ-measure is T-invariant. Moreover ergodicity of $\widetilde{T}$ implies weak mixing. In fact the spectral structure of $U_{\widetilde{T}}$ is very similar to the Gaussian one: namely the first chaos equals $L^2(X, \mathcal{B}, \mu)$ (we emphasize that this is about the whole L^2 and not only L_0^2) on which $U_{\widetilde{T}}$ acts as U_T and the $L^2(\widetilde{X}, \widetilde{\mu})$ together with the action of $U_{\widetilde{T}}$ has the structure of the symmetric Fock space $F(L^2(X, \mathcal{B}, \mu))$ (see the glossary).

We refer to [22,86,168,169] for ergodic properties of systems given by symmetric α-stable stationary processes, or more generally infinitely divisible processes. Again, they share spectral properties similar to the Gaussian case: ergodicity implies weak mixing, while mixing implies mixing of all orders.

In [171], E. Roy clarifies the dynamical "status" of such systems. He uses Poisson suspension automorphisms and the Maruyama representation of an infinitely divisible process mixed with basic properties of automorphisms preserving infinite measure (see [1]) to prove that as a dynamical system, a stationary infinitely divisible process (without the Gaussian part), is a factor of the Poisson suspension over the Lévy measure of this process. In [170] a theory of ID-joinings is developed (which should be viewed as an analog of the GAG theory in the Gaussian class). Parreau and Roy [155] give an example of a Poisson suspension with a minimal possible set of ergodic self-joinings.

Many natural problems still remain open here, for example (assuming always zero entropy of the dynamical system under consideration): Are Poisson suspensions disjoint from Gaussian systems? What is the spectral structure for dynamical systems generated by symmetric α-stable process? Are such systems disjoint whenever $\alpha_1 \neq \alpha_2$? Are Poissonian systems disjoint from local rank one automorphisms (cf. [181])? In [91] it is proved that Gaussian systems are disjoint from so called simple systems (see ▶ Joinings in Ergodic Theory and [93,208]); we will come back to an extension of this result in Sect. "Future Directions". It seems that flows of probabilistic origin satisfy the Kolmogorov group property for the spectrum. One can therefore ask how different are systems satisfying the Kolmogorov group property from systems for which the convolutions of the maximal spectral type are pairwise disjoint (see also Sect. "Future Directions" and the SCS property).

We also mention here another problem which will be taken up in Sect. "Special Flows and Flows on Surfaces, Interval Exchange Transformations" – *Is it true that flows of probabilistic origin are disjoint from smooth flows on surfaces?* Recently A. Katok and A. Windsor announced that it is possible to construct a Kronecker measure so that the corresponding Gaussian system ($\mathbb{Z}$-action (!)) has a smooth representation on the torus.

Yet one more (joining) property seems to be characteristic in the class of systems of probabilistic origin, namely they satisfy so called ELF property (see [30] and ▶ Joinings in Ergodic Theory). Vershik asked whether the ELF property is spectral – however the example of a cocycle from [205] together with Theorem 7 (i) yields a certain Rokhlin extension of a rotation which is ELF and has countable Lebesgue spectrum in the orthocomplement of the eigenfunctions (see [206]); on the other hand any affine extension of that rotation is spectrally the same, while it cannot have the ELF property.

Prikhodko and Thouvenot (private communication) have constructed weakly mixing and non-mixing rank one automorphisms which enjoy the ELF property.

Inducing and Spectral Theory

Assume that T is an ergodic automorphism of a standard probability Borel space $(X, \mathcal{B}, \mu)$. Can "all" dynamics be obtained by inducing (see the glossary) from one fixed automorphism was a natural question from the very beginning of ergodic theory. Because of Abramov's formula for entropy $h(T_A) = h(T)/\mu(A)$ it is clear that positive entropy transformations cannot be obtained from inducing on a zero entropy automorphism. However here we are interested in spectral questions and thus we ask how many spectral types we obtain when we induce. It is proved in [59] that the family of $A \in \mathcal{B}$ for which T_A is mixing is dense for the (pseudo) metric $d(A_1, A_2) = \mu(A_1 \triangle A_2)$. De la Rue [182] proves the following result: *For each ergodic transformation T of a standard probability space $(X, \mathcal{B}, \mu)$ the set of $A \in \mathcal{B}$ for which the maximal spectral type of U_{T_A} is Lebesgue is dense in $\mathcal{B}$.* The multiplicity function is not determined in that paper. Recall (without giving a formal definition, see [152]) that a zero entropy automorphism is *loosely Bernoulli* (LB for short) if and only if it can be induced from an irrational rotation (see also [43,99]). The LB theory shows that not all dynamical systems can be obtained by inducing from an ergodic rotation. However an open question remained whether LB systems exhaust spectrally all Koopman representations. In a deep paper [180], de la Rue studies LB property in the class of Gaussian–Kronecker automorphisms, in particular he constructs S which is not LB. Suppose now that T is LB and for some $A \in \mathcal{B}$, U_{T_A} is isomorphic to U_S. Then by the Foiaş–Stratila Theorem, T_A is isomorphic to S, and hence T_A is not LB. However an induced automorphism from an LB automorphism is LB, a contradiction.

Special Flows and Flows on Surfaces, Interval Exchange Transformations

We now turn our attention to flows. The cases of the geodesic flow, horocycle flows on homogenous spaces of $SL(2, \mathbb{R})$ and nilflows are classical (we refer the reader to [105] with a nice description of the first two cases, while for nilflows we refer to [157]: these classes of flows on homogenous spaces have countable Lebesgue spectrum, in the third case – in the orthocomplement of the eigenspace). On the other hand the classical cyclic approximation theory of Katok and Stepin [104] (see [26]) leads to examples of smooth flows on the torus with simple continuous singular spectra.

Given an ergodic automorphism T on $(X, \mathcal{B}, \mu)$ and a positive integrable function $f: X \to \mathbb{R}^+$ consider the corresponding special flow T^f (see the glossary). Obviously, such a flow is ergodic. Special flows were introduced to ergodic theory by von Neumann in his fundamental work [148] in 1932. Also in that work he explains how to compute eigenvalues for special flows, namely: $r \in \mathbb{R}$ *is an eigenvalue of* T^f *if and only if the following functional equation*

$$e^{2\pi i r f(x)} = \frac{\xi(x)}{\xi(Tx)}$$

has a measurable solution $\xi: X \to \mathbb{T}$. We recall also that the classical Ambrose-Kakutani theorem asserts that practically each ergodic flow has a special representation ([26], see also Rudolph's theorem on special representation therein).

A classical situation when we obtain "natural" special representations is while considering smooth flows on surfaces (we refer the reader to Hasselblatt's and Katok's monograph [73]). They have transversals on which the Poincaré map is piecewise isometric, and this entails a study of interval exchange transformations (IET), see [26,108,163]. Formally, to define IET of m intervals we need a permutation π of $\{1, \ldots, m\}$ and a probability vector $\lambda = (\lambda_1, \ldots, \lambda_m)$ (with positive entries). Then we define $T = T_{\lambda,\pi}$ of $[0, 1)$ by putting

$$T_{\lambda,\pi}(x) = x + \beta_i^\pi - \beta_i \quad \text{for} \quad x \in [\beta_i, \beta_{i+1}) ,$$

where $\beta_i = \sum_{j<i} \lambda_j$, $\beta_i^\pi = \sum_{\pi j < \pi i} \beta_j$. Obviously, each IET preserves Lebesgue measure. One of possible approaches to study ergodic properties of IET is an "a.e" approach "seen" in the definition of $T_{\lambda,\pi}$. It is based on the fundamental fact that the induced transformation on a subinterval of $[0, 1)$ is also IET (see [26]). This leads to a very delicate and deep mathematics based on Rauzy induction, which is a way of inducing on special intervals, considering only irreducible permutations whose set is partitioned into orbits of some maps (any such an orbit is called a *Rauzy class*). If now $\mathcal{R}$ is a Rauzy class of permutations and λ lies in the standard simplex Δ_{m-1} then the Rauzy induction together with a natural renormalization leads to a map $\mathcal{P}: \mathcal{R} \times \Delta_{m-1} \to \mathcal{R} \times \Delta_{m-1}$. For a better understanding of the dynamics of the Rauzy map Veech [209] introduced the space of *zippered rectangles*. A zippered rectangle associated to the Rauzy class $\mathcal{R}$ is a quadruple (λ, h, a, π), where $\lambda \in \mathbb{R}^m_+$, $h \in \mathbb{R}^m_+$, $a \in \mathbb{R}^m_+$, $\pi \in \mathcal{R}$ and the vectors h and a satisfy some equations and inequalities. Every zippered rectangle (λ, h, a, π) determines a Riemann structure on a compact connected surface. Denote by $\Omega(\mathcal{R})$ the space of all zippered rectangles, corresponding to a given Rauzy class $\mathcal{R}$ and satisfying the condition $\langle \lambda, h \rangle = 1$. In [209], Veech defined a flow $(P^t)_{t\in\mathbb{R}}$ on the space $\Omega(\mathcal{R})$ putting

$$P^t(\lambda, h, a, \pi) = (e^t \lambda, e^{-t} h, e^{-t} a, \pi)$$

and extended the Rauzy map. On so called *Veech moduli space* of zippered rectangles, the flow (P^t) becomes the *Teichmüller flow* and it preserves a natural Lebesgue measure class; by passing to a transversal and projecting the measure on the space of IETs $\mathcal{R} \times \Delta_{m-1}$ Veech has proved the following fundamental theorem ([209], see also [143]) which is a generalization of the fact that Gauss measure $1/(\ln 2) 1/(1 + x) dx$ is invariant for the Gauss map which sends $t \in (0, 1)$ into the fractional part of its inverse.

Theorem 10 (Veech, Masur, 1982) *Assume that* $\mathcal{R}$ *is a Rauzy class. There exists a* σ*-finite measure* $\mu_\mathcal{R}$ *on* $\mathcal{R} \times \Delta_{m-1}$ *which is* $\mathcal{P}$*-invariant, equivalent to "Lebesgue" measure, conservative and ergodic.*

In [209] it is proved that a.e. (in the above sense) IET is then of rank one (and hence is ergodic and has a simple spectrum). He also formulated as an open problem whether we can achieve the weak mixing property a.e. This has been recently answered in positive by A. Avila and G. Forni [19] (for π which is not a rotation).

Katok [100] proved that IET have no mixing factors (in fact his proof shows more: the IET class is disjoint with the class of mixing transformations). By their nature, IET transformations are of finite rank (see [26]) so they are of finite multiplicity. They need not be of simple spectrum (see remarks in [105] pp. 88–90). It remains an open question whether an IET can have a non-singular spectrum. Joining properties in the class of exchange of 3 and more intervals are studied in [47,48]. An important question of Veech [208] whether a.e. IET is simple is still open.

When we consider a smooth flow on a surface preserving a smooth measure, whose only singularity (we assume that we have only finitely many singularities) are simple (non-degenerated) saddles then such a flow has a special representation over an interval exchange automorphism under a smooth function which has finitely many logarithmic singularities (see [73]). In the article by Arnold [18] the quasi-periodic Hamiltonian case is considered: $H: \mathbb{R}^2 \to \mathbb{R}$ satisfies $H(x+m, y+n) = H(x, y) + n\alpha_1 + m\alpha_2$, $\alpha_1/\alpha_2 \notin \mathbb{Q}$, and we then consider the following system of differential equations on $\mathbb{T}^2$

$$\frac{dx}{dt} = \frac{\partial H}{\partial y}, \quad \frac{dy}{dt} = \frac{-\partial H}{\partial x} . \tag{2}$$

As Arnold shows the dynamical system arising from the system (2) has one ergodic component which has a special representation over the irrational rotation by $\alpha := \alpha_1/\alpha_2$ under a smooth function with finitely many logarithmic singularities (all other ergodic components are periodic orbits and separatrices). By changing a speed as it is done in [57] so that critical points of the vector field in (2) become singular points, Arnold's special representation is transformed to a special flow over the same irrational rotation however under a piecewise smooth function. If the sum of jumps is not zero then in fact we come back to von Neumann's class of special flows considered in [148]. Similar classes of special flows (when the roof function is of bounded variation) are obtained from ergodic components of flows associated to billiards in convex polygones with rational angles [106]. Kochergin [115] showed that special flows over irrational rotations and under bounded variation functions are never mixing. This has been recently strengthened in [54] to the following: *If T is an irrational rotation and f is of bounded variation then the special flow T^f is spectrally disjoint from all mixing flows*. In particular all such flows have singular spectra. Moreover, in [54] it is proved that whenever the Fourier transform of the roof function f is of order $O(1/n)$ then T^f is disjoint from all mixing flows (see also [55]). In fact in the papers [54,55,56,57] the authors discuss the problem of disjointness of those special flows with all ELF-flows conjecturing that no flow of probabilistic origin has a smooth realization on a surface. In [140] the analytic case is considered leading to a "generic" result on disjointness with the ELF class generalizing the classical Shklover's result on the weak mixing property [197].

Kochergin [117] proved the absence of mixing for flows where the roof function has finitely many singularities, whenever the sum of "left logarithmic speeds" and the sum of "right logarithmic speeds" are equal – this is called a *symmetric logarithmic case*, however some Diophantine restriction is put on α.

In [128], where also the absence of mixing is considered for the symmetric logarithmic case, it was conjectured (and proved for arbitrary rotation) that a necessary condition for mixing of a special flow T^f (with arbitrary T and f) is the condition that the sequence of distributions $((f_0^{(n)})_*)_n$ tends to δ_∞ in the space of probability measures on $\overline{\mathbb{R}}$. K. Schmidt [194] proved it using the theory of cocycles and extending a result from [3] on tightness of cocycles.

A. Katok [100] proved the absence of mixing for special flows over IET when the roof function is of bounded variation (see also [187]). Katok's theorem was strengthened in [56] to the disjointness theorem with the class of mixing flows.

On the other hand there is a lot of (difficult) results pointing out classes of special flows over irrational rotations which are mixing, especially (but not only) in the class of non-symmetric logarithmic singularities: [36,38] (B. Fayad was able to give a speed of convergence to zero for Fourier coefficients), [109,119,120]. Recently mixing property has been proved in a non-symmetric case in [203] when the base transformation is a special class of IETs.

The eigenvalue problem (mainly how many frequencies can have the group of eigenvalues) for special flows over irrational rotations is studied in [41,42,71].

A. Avila and G. Forni [19] proved that a.e. translation flow on a surface (of genus at least two) is weakly mixing (which is a drastic difference with the linear flow case of the torus, where the spectrum is always discrete).

The problem of whether mixing flows indicated in this chapter are mixing of all orders is open (it is also unknown whether they have singular spectra). One of several possible approaches (proposed by B. Fayad and J.-P. Thouvenot) toward positive solution of this problem would be to show that such flows enjoy so called Ratner's property (R-property). This property may be viewed as a particular way of divergence of orbits of close points; it was shown to hold for horocycle flows by M. Ratner [162]. We refer the reader to [162] and the survey article [201] for the formal definitions and basic consequences of R-property. In particular, R-property implies "rigidity" of joinings and it also implies the PID property; hence mixing and R-property imply mixing of all orders. In [57,58] a version of R-property is shown for the class of von Neumann special flows (however α is assumed to have bounded partial quotients). This allowed one to prove there that such flows are even mildly mixing (mixing is excluded by a Kochergin's result). We conjecture that an R-property may also hold for special flows over multidimensional rotations with roof functions given by nil-cocycles which we mentioned in Sect. "Spectral Theory of Weighted Operators".

If indeed the R-property is ubiquitous in the class of smooth flows on surfaces it may also be useful to show that smooth flows on surfaces are disjoint with flows of probabilistic origin – see [91,92,135,190,202].

B. Fayad [40] gives a criterion that implies singularity of the maximal spectral type for a dynamical system on a Riemannian manifold. As an application he gives a class of smooth mixing flows (with singular spectra) on $\mathbb{T}^3$ obtained from linear flows by a time change (again this is a drastic difference with dimension two, where a smooth time change of a linear flow leads to non-mixing flows [26]).

The spectral multiplicity problem for special flows (with sufficiently regular roof functions) over irrational rotations seems to be completely untouched (except for the case of a sufficiently smooth f – the spectrum of T^f is then simple [26]). It would be nice to have examples of such flows with finite bigger than one multiplicity. In particular, is it true that the von Neumann class of special flows have finite multiplicity? This was partially solved by A. Katok (private communication) on certain subspaces in L^2, but not on the whole L^2-space.

Problem. *Given $Tx = x + \alpha$ (with α irrational) can we find $f\colon [0,1) \to \mathbb{R}^+$ sufficiently regular (e. g. with finitely many "controllable" singularities) such that T^f has a Lebesgue spectrum?*

Of course the above is related to the question whether at all one can find a smooth flow on a surface with a Lebesgue spectrum (for $\mathbb{Z}$-actions we can even see positive entropy diffeomorphisms on the torus).

We mention at the end that if we drop here (and in other problems) the assumption of regularity of f then the answers will be always positive because of the LB theory; in particular there is a section of any horocycle flow (it has the LB property [161]) such that in the corresponding special representation T^f the map T is an irrational rotation. Using a Kochergin's result [118] on cohomology (see also [98,176]) the L^1-function f is cohomologous to a positive function g which is even continuous, thus T^f is isomorphic to T^g.

Future Directions

We have already seen several cases where spectral properties interact with measure-theoretic properties of a system. Let us mention a few more cases which require further research and deeper understanding.

We recall that the weak mixing property can be understood as a property complementary to discrete spectrum (more precisely to the distality [62]), or similarly mild mixing property is complementary to rigidity. This can be phrased quite precisely by saying that T is not weakly (mildly) mixing if and only if it has a non-trivial factor with discrete spectrum (it has a non-trivial rigid factor). It has been a question for quite a long time if in a sense mixing can be "built" on the same principle. In other words we seek a certain "highly" non-mixing factor. It was quite surprising when in 2005 F. Parrreau (private communication) gave the positive answer to this problem.

Theorem 11 (Parreau) *Assume that T is an ergodic automorphism of a standard probability space $(X, \mathcal{B}, \mu)$. Assume moreover that T is not mixing. Then there exists a non-trivial factor (see below) of T which is disjoint with all mixing automorphisms.*

In fact, Parreau proved that each factor of T given by $\mathcal{B}_\infty(\rho)$ (this σ-algebra is described in [136]), where $U_T^{n_k} \to J_\rho$, is disjoint from all mixing transformations. This proof leads to some other results of the same type, for example: *Assume that T is an ergodic automorphism of a standard probability space. Assume that there exists a non-trivial automorphism S with a singular spectrum which is not disjoint with T. Then T has a non-trivial factor which is disjoint with any automorphism with a Lebesgue spectrum.*

The problem of spectral multiplicity of Cartesian products for "typical" transformation studied by Katok [98] and then its solution in [13] which we already considered in Sect. "The Multiplicity Function" lead to a study of those T for which

$$(CS) \quad \sigma^{(m)} \perp \sigma^{(n)} \text{ whenever } m \neq n\,,$$

where $\sigma = \sigma_T$ just stands for the reduced maximal spectral type of U_T (which is constantly assumed to be a continuous measure), see also Stepin's article [199].

The usefulness of the above property (CS) in ergodic theory was already shown in [90], where a spectral counterexample machinery was presented using the following observation: *If $\mathcal{A}$ is a $T^{\times\infty}$-invariant sub-σ-algebra such that the maximal spectral type on $L^2(\mathcal{A})$ is absolutely continuous with respect to σ_T then $\mathcal{A}$ is contained in one of the coordinate sub-σ-algebras $\mathcal{B}$.* Based on that in [90] it is shown how to construct two weakly isomorphic action which are not isomorphic or how to construct two non-disjoint automorphisms which have no common non-trivial factors (such constructions were previously known for so called minimal self-joining automorphisms [174]). See also [200] for extensions of those results to $\mathbb{Z}^d$-actions.

Prikhodko and Ryzhikov [159] proved that the classical Chacon transformation enjoys the (CS) property. The SCS property defined in the glossary is stronger than the (CS) condition above; the SCS property implies that the corresponding Gaussian system S_{σ_T} has a simple spectrum. Ageev [10] shows that Chacon's transformation satisfies the SCS property; moreover in [13] he shows that the SCS property is satisfied generically and he gives a construction of a rank one mixing SCS-system (see also [191]). In [134] it is proved that some special flows considered in Sect. "Special Flows and Flows on Surfaces, Interval Exchange Transformations" (including the von Neumann class, however with α having unbounded partial quotients) have the SCS property. Since the corresponding Gaussian systems have simple spectra, it would be interesting to decide whether σ_T (for an SCS-automorphism) can be concentrated on a set without rational relations. It is quite plausible that the SCS property is commonly seen for smooth flows on surfaces.

Katok and Thouvenot (private communication) considered systems called *infinitely divisible*. These are systems T on $(X, \mathcal{B}, \mu)$ which have a family of factors $\mathcal{B}_\omega$ indexed by $\omega \in \bigcup_{n=0}^{\infty}\{0,1\}^n$ ($\mathcal{B}_\varepsilon = \mathcal{B}$) such that $\mathcal{B}_{\omega 0} \perp \mathcal{B}_{\omega 1}$, $\mathcal{B}_{\omega 0} \vee \mathcal{B}_{\omega 1} = \mathcal{B}_\omega$ and for each $\eta \in \{0,1\}^{\mathbb{N}}$, $\bigcap_{n\in\mathbb{N}} \mathcal{B}_{\eta[0,n]} = \{\emptyset, X\}$. They showed (unpublished) that there are discrete spectrum transformations which are ID, and that there are rank one transformations with continuous spectra which are also ID (clearly Gaussian systems are ID). It was until recently that a relationship between ID automorphisms and systems coming from stationary ID processes was unclear. In [135] it is proved that dynamical systems coming from stationary ID processes are factors of ID automorphisms; moreover, ID automorphisms are disjoint with all systems having the SCS property. It would be nice to decide whether Koopman representations associated to ID automorphisms satisfy the Kolmogorov group property.

Acknowledgments

Research supported by the EU Program Transfer of Knowledge "Operator Theory Methods for Differential Equations" TODEQ and the Polish Ministry of Science and High Education.

Bibliography

1. Aaronson J (1997) An introduction to infinite ergodic theory, mathematical surveys and monographs. Am Math Soc 50:284
2. Aaronson J, Nadkarni MG (1987) L^∞ eigenvalues and L^2 spectra of nonsingular transformations. Proc London Math Soc 55(3):538–570
3. Aaronson J, Weiss B (2000) Remarks on the tightness of cocycles. Coll Math 84/85:363–376
4. Abdalaoui H (2000) On the spectrum of the powers of Ornstein transformations Ergodic theory and harmonic analysis (Mumbai, 1999). Sankhya Ser. A 62:291–306
5. Abdalaoui H (2007) A new class of rank 1 transformations with singular spectrum. Ergodic Theor Dynam Syst 27:1541–1555
6. Abdalaoui H, Parreau F, Prikhodko A (2006) A new class of Ornstein transformations with singular spectrum. Ann Inst H Poincaré Probab Statist 42:671–681
7. Adams T (1998) Smorodinsky's conjecture on rank one systems. Proc Am Math Soc 126:739–744
8. Ageev ON (1988) Dynamical systems with a Lebesgue component of even multiplicity in the spectrum. Mat Sb (NS) 136(178):307–319, (Russian) 430 (1989) Translation in Math. USSR-Sb. 64:305–317
9. Ageev ON (1999) On ergodic transformations with homogeneous spectrum. J Dynam Control Syst 5:149–152
10. Ageev ON (2000) On the spectrum of cartesian powers of classical automorphisms. Mat Zametki 68:643–647, (2000) Translation Math. Notes 68:547–551, (Russian)
11. Ageev ON (2001) On the multiplicity function of generic group extensions with continuous spectrum. Ergodic Theory Dynam Systems 21:321–338
12. Ageev ON (2005) The homogeneous spectrum problem in ergodic theory. Invent Math 160:417–446
13. Ageev ON (2007) Mixing with staircase multiplicity function. preprint
14. Alexeyev VM (1958) Existence of bounded function of the maximal spectral type. Vestnik Mosc Univ 5:13-15 and (1982) Ergodic Theory Dynam Syst 2:259–261
15. Anosov DV (2003) On Spectral multiplicities in ergodic theory. Institut im. V.A. Steklova, Moscow
16. Anzai H (1951) Ergodic skew product transformations on the torus. Osaka J Math 3:83–99
17. Arni N (2005) Spectral and mixing properties of actions of amenable groups. Electron Res Announc AMS 11:57–63
18. Arnold VI (1991) Topological and ergodic properties of closed 1-forms with incommensurable periods. Funktsional Anal Prilozhen 25:1–12, (Russian)
19. Avila A, Forni G (2007) Weak mixing for interval exchange transformations and translation flows. Ann Math 165:637–664
20. Baxter JR (1971) A class of ergodic transformations having simple spectrum. Proc Am Math Soc 27:275–279
21. Bourgain J (1993) On the spectral type of Ornstein's class of transformations. Isr J Math 84:53–63
22. Cambanis S, Podgórski K, Weron A (1995) Chaotic behaviour of infinitely divisible processes. Studia Math 115:109–127
23. Chacon RV (1970) Approximation and spectral multiplicity. In: Dold A, Eckmann B (eds) Contributions to ergodic theory and probability. Springer, Berlin, pp 18–27
24. Choe GH Products of operators with singular continuous spectra. In: Operator theory: operator algebras and applications, Part 2. (Durham, 1988), pp 65–68, Proc Sympos Pure Math, 51, Part 2, Amer Math Soc, Providence, 1990
25. Connes A, Woods E (1985) Approximately transitive flows and ITPFI factors. Ergod Theory Dynam Syst 5:203–236
26. Cornfeld IP, Fomin SV, Sinai YG (1982) Ergodic theory. Springer, New York
27. Danilenko AL (2006) Explicit solution of Rokhlin's problem on homogeneous spectrum and applications. Ergod Theory Dyn Syst 26:1467–1490
28. Danilenko AL, Lemańczyk M (2005) A class of multipliers for $\mathcal{W}^\perp$. Isr J Math 148:137–168
29. Danilenko AI, Park KK (2002) Generators and Bernouillian factors for amenable actions and cocycles on their orbits. Ergod Theory Dyn Syst 22:1715–1745
30. Derriennic Y, Frączek K, Lemańczyk M, Parreau F (2007) Ergodic automorphisms whose weak closure of off-diagonal measures consists of ergodic self-joinings. Coll Math 110:81–115
31. Dooley A, Golodets VY (2002) The spectrum of completely positive entropy actions of countable amenable groups. J Funct Anal 196:1–18
32. Dooley A, Quas A (2005) Approximate transitivity for zero-entropy systems. Ergod Theory Dyn Syst 25:443–453

33. Downarowicz T, Kwiatkowski J (2000) Spectral isomorphism of Morse flows. Fundam Math 163:193–213
34. Downarowicz T, Kwiatkowski J (2002) Weak Cosure Theorem fails for $\mathbb{Z}^d$-actions. Stud Math 153:115–125
35. Fayad B (2001) Partially mixing and locally rank 1 smooth transformations and flows on the torus $\mathbb{T}^d, d \geq 3$. J London Math Soc 64(2):637–654
36. Fayad B (2001) Polynomial decay of correlations for a class of smooth flows on the two torus. Bull Soc Math France 129:487–503
37. Fayad B (2002) Skew products over translations on $T^d, d \geq 2$. Proc Amer Math Soc 130:103–109
38. Fayad B (2002) Analytic mixing reparametrizations of irrational flows. Ergod Theory Dyn Syst 22:437–468
39. Fayad B (2005) Rank one and mixing differentiable flows. Invent Math 160:305–340
40. Fayad B (2006) Smooth mixing flows with purely singular spectra. Duke Math J 132:371–391
41. Fayad B, Katok AB, Windsor A (2001) Mixed spectrum reparameterizations of linear flows on $\mathbb{T}^2$, dedicated to the memory of I. G. Petrovskii on the occasion of his 100th anniversary. Mosc Math J 1:521–537
42. Fayad B, Windsor A (2007) A dichotomy between discrete and continuous spectrum for a class of special flows over rotations. J Mod Dyn 1:107–122
43. Feldman J (1976) New K-automorphisms and a problem of Kakutani. Isr J Math 24:16–38
44. Ferenczi S (1984) Systèmes localement de rang un. Ann Inst H Poincaré Probab Statist 20:35–51
45. Ferenczi S (1985) Systèmes de rang un gauche. Ann Inst H Poincaré Probab Statist 21:177–186
46. Ferenczi S (1997) Systems of finite rank. Colloq Math 73:35–65
47. Ferenczi S, Holton C, Zamboni LQ (2005) Joinings of three-interval exchange transformations. Ergod Theory Dyn Syst 25:483–502
48. Ferenczi S, Holton C, Zamboni LQ (2004) Structure of three-interval exchange transformations III: ergodic and spectral properties. J Anal Math 93:103–138
49. Ferenczi S, Lemańczyk M (1991) Rank is not a spectral invariant. Stud Math 98:227–230
50. Filipowicz I (1997) Product $\mathbb{Z}^d$-actions on a Lebesgue space and their applications. Stud Math 122:289–298
51. Foiaş C, Stratila S (1968) Ensembles de Kronecker dans la théorie ergodique. CR Acad Sci Paris, Ser A-B 267:A166–A168
52. Frączek K (1997) On a function that realizes the maximal spectral type. Stud Math 124:1–7
53. Frączek K (2000) Circle extensions of Z^d-rotations on the d-dimensional torus. J London Math Soc 61(2):139–162
54. Frączek K, Lemańczyk M (2004) A class of special flows over irrational rotations which is disjoint from mixing flows. Ergod Theory Dyn Syst 24:1083–1095
55. Frączek K, Lemańczyk M (2003) On symmetric logarithm and some old examples in smooth ergodic theory. Fund Math 180:241–255
56. Frączek K, Lemańczyk M (2005) On disjointness properties of some smooth flows. Fund Math 185:117–142
57. Frączek K, Lemańczyk M (2006) On mild mixing of special flows over irrational rotations under piecewise smooth functions. Ergod Theory Dyn Syst 26:719–738
58. Frączek K, Lemańczyk M, Lesigne E (2007) Mild mixing property for special flows under piecewise constant functions. Disc Contin Dyn Syst 19:691–710
59. Friedman NA, Ornstein DS (1973) Ergodic transformations induce mixing transformations. Adv Math 10:147–163
60. Furstenberg H (1967) Disjointness in ergodic theory, minimal sets, and a problem of Diophantine approximation. Math Syst Theory 1:1–49
61. Furstenberg H, Weiss B (1978) The finite multipliers of infinite ergodic transformations. Lect Notes Math 668:127–132
62. Furstenberg H (1981) Recurrence in ergodic theory and combinatorial number theory. Princeton University Press, Princeton
63. Glasner E (1994) On the class of multipliers for $\mathcal{W}^{\perp}$. Ergod Theory Dyn Syst 14:129–140
64. Glasner E (2003) Ergodic theory via joinings. Math Surv Monogr, AMS, Providence, RI 101:384
65. Glasner E, Weiss B (1989) Processes disjoint from weak mixing. Trans Amer Math Soc 316:689–703
66. Goodson GR (1999) A survey of recent results in the spectral theory of ergodic dynamical systems. J Dyn Control Syst 5:173–226
67. Goodson GR, Kwiatkowski J, Lemańczyk M, Liardet P (1992) On the multiplicity function of ergodic group extensions of rotations. Stud Math 102:157–174
68. Gromov AL (1991) Spectral classification of some types of unitary weighted shift operators. Algebra Anal 3:62–87, (Russian) (1992) Translation in St. Petersburg Math J 3:997–1021
69. Guenais M (1998) Une majoration de la multiplicité spectrale d'opérateurs associés à des cocycles réguliers. Isr J Math 105:263–284
70. Guenais M (1999) Morse cocycles and simple Lebesgue spectrum. Ergod Theory Dyn Syst 19:437–446
71. Guenais M, Parreau F (2005) Valeurs propres de transformations liées aux rotations irrationnelles et aux fonctions en escalier. (in press)
72. Hahn F, Parry W (1968) Some characteristic properties of dynamical systems with quasi-discrete spectra. Math Syst Theory 2:179–190
73. Hasselblatt B, Katok AB (2002) Principal structures. In: Handbook of dynamical systems, vol 1A. North-Holland, Amsterdam, pp 1–203
74. Helson H (1986) Cocycles on the circle. J Oper Theory 16:189–199
75. Helson H, Parry W (1978) Cocycles and spectra. Arkiv Math 16:195–206
76. Herman M (1979) Sur la conjugaison différentiable des difféomorphismes du cercle à des rotations. Publ Math IHES 49:5–234
77. Host B (1991) Mixing of all orders and pairwise independent joinings of systems with singular spectrum. Isr J Math 76:289–298
78. Host B, Méla F, Parreau F (1991) Non-singular transformations and spectral anaysis of measures. Bull Soc Math France 119:33–90
79. Iwanik A (1991) The problem of L^p-simple spectrum for ergodic group automorphisms. Bull Soc Math France 119:91–96
80. Iwanik A (1992) Positive entropy implies infinite L^p-multiplicity for $p > 1$. Ergod Theory Rel Top (Güstrow 1990) III:124–127. Lecture Notes in Math, vol 1514. Springer, Berlin

81. Iwanik A (1997) Anzai skew products with Lebesgue component of infinite multiplicity. Bull London Math Soc 29:195–199
82. Iwanik A, Lemańczyk M, Rudolph D (1993) Absolutely continuous cocycles over irrational rotations. Isr J Math 83:73–95
83. Iwanik A, Lemańczyk M, de Sam Lazaro J, de la Rue T (1997) Quelques remarques sur les facteurs des systèmes dynamiques gaussiens. Stud Math 125:247–254
84. Iwanik A, Lemańczyk M, Mauduit C (1999) Piecewise absolutely continuous cocycles over irrational rotations. J London Math Soc 59(2):171–187
85. Iwanik A, de Sam Lazaro J (1991) Sur la multiplicité L^p d'un automorphisme gaussien. CR Acad Sci Paris Ser I 312:875–876
86. Janicki A, Weron A (1994) Simulation and chaotic behavior of α-stable stochastic processes. Monographs and Textbooks in Pure and Applied Mathematics, vol 178. Marcel Dekker, New York, pp 355
87. del Junco A (1976) Transformations with discrete spectrum are stacking transformations. Canad J Math 28:836–839
88. del Junco A (1977) A transformation with simple spectrum which is not of rank one. Canad J Math 29:655–663
89. del Junco A (1981) Disjointness of measure-preserving transformations, minimal self-joinings and category. Prog Math 10:81–89
90. del Junco A, Lemańczyk M (1992) Generic spectral properties of measure-preserving maps and applications. Proc Amer Math Soc 115:725–736
91. del Junco A, Lemańczyk M (1999) Simple systems are disjoint with Gaussian systems. Stud Math 133:249–256
92. del Junco A, Lemańczyk M (2005) Joinings of distally simple systems. (in press)
93. del Junco A, Rudolph D (1987) On ergodic actions whose self-joinings are graphs. Ergod Theory Dyn Syst 7:531–557
94. Kachurovskii AG (1990) A property of operators generated by ergodic automorphisms. Optimizatsiya 47(64):122–125, (Russian)
95. Kalikow S (1984) Two fold mixing implies three fold mixing for rank one tranformations. Ergod Theory Dyn Syst 4:237–259
96. Kamae T Spectral properties of automata generating sequences. (in press)
97. Kamiński B (1981) The theory of invariant partitions for $\mathbb{Z}^d$-actions. Bull Acad Polon Sci Ser Sci Math 29:349–362
98. Katok AB Constructions in ergodic theory. (in press)
99. Katok AB (1977) Monotone equivalence in ergodic theory. Izv Akad Nauk SSSR Ser Mat 41:104–157, (Russian)
100. Katok AB (1980) Interval exchange transformations and some special flows are not mixing. Isr J Math 35:301–310
101. Katok AB (2001) Cocycles, cohomology and combinatorial constructions in ergodic theory. In: Robinson Jr. EA (ed) Proc Sympos Pure Math 69, Smooth ergodic theory and its applications, Seattle, 1999. Amer Math Soc, Providence, pp 107–173
102. Katok AB (2003) Combinatorial constructions in ergodic theory and dynamics. In: University Lecture Series 30. Amer Math Soc, Providence
103. Katok AB, Lemańczyk M (2007) Some new cases of realization of spectral multiplicity function for ergodic transformations. (in press)
104. Katok AB, Stepin AM (1967) Approximations in ergodic theory. Uspekhi Mat Nauk 22(137):81–106, (Russian)
105. Katok AB, Thouvenot J-P (2006) Spectral Properties and Combinatorial Constructions in Ergodic Theory. In: Handbook of dynamical systems, vol 1B. Elsevier, Amsterdam, pp 649–743
106. Katok AB, Zemlyakov AN (1975) Topological transitivity of billiards in polygons. Mat Zametki 18:291–300, (Russian)
107. Keane M (1968) Generalized Morse sequences. Z Wahr Verw Geb 10:335–353
108. Keane M (1975) Interval exchange transformations. Math Z 141:25–31
109. Khanin KM, Sinai YG (1992) Mixing of some classes of special flows over rotations of the circle. Funct Anal Appl 26:155–169
110. King JL (1986) The commutant is the weak closure of the powers, for rank-1 transformations. Ergod Theory Dyn Syst 6:363–384
111. King JL (1988) Joining-rank and the structure of finite rank rank mixing transformations. J Anal Math 51:182–227
112. King JL (2001) Flat stacks, joining-closure and genericity. (in press)
113. Klemes I (1996) The spectral type of the staircase transformation. Tohoku Math J 48:247–248
114. Klemes I, Reinhold K (1997) Rank one transformations with singular spectral type. Isr J Math 98:1–14
115. Kocergin AV (1972) On the absence of mixing in special flows over the rotation of a circle and in flows on a two-dimensional torus. Dokl Akad Nauk SSSR 205:949–952, (Russian)
116. Kochergin AV (1975) Mixing in special flows over rearrangment of segments and in smooth flows on surfaces. Math USSR Acad Sci 25:441–469
117. Kochergin AV (1976) Non-degenerate saddles and absence of mixing. Mat Zametky 19:453–468, (Russian)
118. Kochergin AV (1976) On the homology of function over dynamical systems. Dokl Akad Nauk SSSR 231:795–798
119. Kochergin AV (2002) A mixing special flow over a rotation of the circle with an almost Lipschitz function. Sb Math 193:359–385
120. Kochergin AV (2004) Nondegenerate fixed points and mixing in flows on a two-dimensional torus II. Math Sb 195:15–46, (Russian)
121. Koopman BO (1931) Hamiltonian systems and transformations in Hilbert space. Proc Nat Acad Sci USA 17:315–318
122. Kuipers L, Niederreiter H (1974) Uniform distribution of sequences. Wiley, London, pp 407
123. Kushnirenko AG (1974) Spectral properties of some dynamical systems with polynomial divergence of orbits. Vestnik Moskovskogo Univ 1–3:101–108
124. Kwiatkowski J (1981) Spectral isomorphism of Morse dynamical systems Bull Acad Polon Sci Ser Sci Math 29:105–114
125. Kwiatkowski Jr. J, Lemańczyk M (1995) On the multiplicity function of ergodic group extensions II. Stud Math 116:207–215
126. Lemańczyk M (1988) Toeplitz Z_2–extensions. Ann Inst H Poincaré 24:1–43
127. Lemańczyk M (1996) Introduction to ergodic theory from the point of view of the spectral theory. In: Lecture Notes of the tenth KAIST mathematics workshop, Taejon, 1996, pp 1–153
128. Lemańczyk M (2000) Sur l'absence de mélange pour des flots spéciaux au dessus d'une rotation irrationnelle. Coll Math 84/85:29–41
129. Lemańczyk M, Lesigne E (2001) Ergodicity of Rokhlin cocycles. J Anal Math 85:43–86

130. Lemańczyk M, Mauduit C (1994) Ergodicity of a class of cocycles over irrational rotations. J London Math Soc 49:124–132
131. Lemańczyk M, Mentzen MK, Nakada H (2003) Semisimple extensions of irrational rotations. Stud Math 156:31–57
132. Lemańczyk M, Parreau F (2003) Rokhlin extensions and lifting disjointness. Ergod Theory Dyn Syst 23:1525–1550
133. Lemańczyk M, Parreau F (2004) Lifting mixing properties by Rokhlin cocycles. (in press)
134. Lemańczyk M, Parreau F (2007) Special flows over irrational rotation with simple convolution property. (in press)
135. Lemańczyk M, Parreau F, Roy E (2007) Systems with simple convolutions, distal simplicity and disjointness with infinitely divisible systems. (in press)
136. Lemańczyk M, Parreau F, Thouvenot J-P (2000) Gaussian automorphisms whose ergodic self–joinings are Gaussian. Fundam Math 164:253–293
137. Lemańczyk M, de Sam Lazaro J (1997) Spectral analysis of certain compact factors for Gaussian dynamical systems. Isr J Math 98:307–328
138. Lemańczyk M, Sikorski A (1987) A class of not local rank one automorphisms arising from continuous substitutions. Prob Theory Rel Fields 76:421–428
139. Lemańczyk M, Wasieczko M (2006) A new proof of Alexeyev's theorem. (in press)
140. Lemańczyk M, Wysokińska M (2007) On analytic flows on the torus which are disjoint from systems of probabilistic origin. Fundam Math 195:97–124
141. Leonov VP (1960) The use of the characteristic functional and semi–invariants in the ergodic theory of stationary processes. Dokl Akad Nauk SSSR 133:523–526, (1960) Sov Math 1:878–881
142. Maruyama G (1970) Infinitely divisible processes. Theory Prob Appl 15(1):1–22
143. Masur H (1982) Interval exchange transformations and measured foliations. Ann Math 115:169–200
144. Mathew J, Nadkarni MG (1984) Measure-preserving transformation whose spectrum has Lebesgue component of multiplicity two. Bull London Math Soc 16:402–406
145. Medina H (1994) Spectral types of unitary operators arising from irrational rotations on the circle group. Michigan Math J 41:39–49
146. Mentzen MK (1988) Some examples of automorphisms with rank *r* and simple spectrum. Bull Pol Acad Sci 7–8:417–424
147. Nadkarni MG (1998) Spectral theory of dynamical systems. Hindustan Book Agency, New Delhi
148. von Neumann J (1932) Zur Operatorenmethode in der Klassichen Mechanik. Ann Math 33:587–642
149. Newton D (1966) On Gaussian processes with simple spectrum. Z Wahrscheinlichkeitstheorie Verw Geb 5:207–209
150. Newton D, Parry W (1966) On a factor automorphism of a normal dynamical system. Ann Math Statist 37:1528–1533
151. Ornstein D (1970) On the root problem in ergodic theory. In: Proc. 6th Berkeley Symp Math Stats Prob, University California Press, Berkeley, 1970, pp 348–356
152. Ornstein D, Rudolph D, Weiss B (1982) Equivalence of measure preserving transformations. Mem Amer Math Soc 37(262):116
153. Ornstein D, Weiss B (1987) Entropy and isomorphism theorems for actions of amenable groups. J Anal Math 48:1–141
154. Parreau F (2000) On the Foiaş and Stratila theorem. In: Proceedings of the conference on Ergodic Theory, Toruń, 2000
155. Parreau F, Roy E (2007) Poissson suspensions with a minimal set od self-joinings. (in press)
156. Parry W (1981) Topics in ergodic theory. Cambridge Tracts in Mathematics, 75. Cambridge University Press, Cambridge-New York
157. Parry W (1970) Spectral analysis of *G*-extensions of dynamical systems. Topology 9:217–224
158. Petersen K (1983) Ergodic theory. Cambridge University Press, Cambridge
159. Prikhodko AA, Ryzhikov VV (2000) Disjointness of the convolutions for Chacon's automorphism. Dedicated to the memory of Anzelm Iwanik. Colloq Math 84/85:67–74
160. Queffelec M (1988) Substitution dynamical systems – spectral analysis. In: Lecture Notes in Math 1294. Springer, Berlin, pp 240
161. Ratner M (1978) Horocycle flows are loosely Bernoulli. Isr J Math. 31:122–132
162. Ratner M (1983) Horocycle flows, joinings and rigidity of products. Ann Math 118:277–313
163. Rauzy G (1979) Echanges d'intervalles et transformations induites. Acta Arith 34:315–328
164. Robinson EA Jr (1983) Ergodic measure preserving transformations with arbitrary finite spectral multiplicities. Invent Math 72:299–314
165. Robinson EA Jr (1986) Transformations with highly nonhomogeneous spectrum of finite multiplicity. Isr J Math 56:75–88
166. Robinson EA Jr (1988) Nonabelian extensions have nonsimple spectrum. Compos Math 65:155–170
167. Robinson EA Jr (1992) A general condition for lifting theorems. Trans Amer Math Soc 330:725–755
168. Rosiński J, Zak T (1996) Simple condition for mixing of infinitely divisible processes. Stoch Process Appl 61:277–288
169. Rosiński J, Zak T (1997) The equivalence of ergodicity and weak mixing for infinitely divisible processes. J Theor Probab 10:73–86
170. Roy E (2005) Mesures de Poisson, infinie divisibilité et propriétés ergodiques. In: Thèse de doctorat de l'Université Paris 6
171. Roy E (2007) Ergodic properties of Poissonian ID processes. Ann Probab 35:551–576
172. Royden HL (1968) Real analysis. McMillan, New York
173. Rudin W (1962) Fourier analysis on groups. In: Interscience Tracts in Pure and Applied Mathematics, No. 12 Interscience Publishers. Wiley, New York, London, pp 285
174. Rudolph DJ (1979) An example of a measure-preserving map with minimal self-joinings and applications. J Anal Math 35:97–122
175. Rudolph D (1985) *k*-fold mixing lifts to weakly mixing isometric extensions. Ergod Theor Dyn Syst 5:445–447
176. Rudolph D (1986) $\mathbb{Z}^n$ and $\mathbb{R}^n$ cocycle extension and complementary algebras. Ergod Theor Dyn Syst 6:583–599
177. Rudolph D (1990) Fundamentals of measurable dynamics. Oxford Sci Publ, pp 169
178. Rudolph D (2004) Pointwise and L^1 mixing relative to a subsigma algebra. Illinois J Math 48:505–517
179. Rudolph D, Weiss B (2000) Entropy and mixing for amenable group actions. Ann Math 151(2):1119–1150
180. de la Rue T (1996) Systèmes dynamiques gaussiens d'entropie nulle, lâchement et non lâchement Bernoulli. Ergod Theor Dyn Syst 16:379–404

181. de la Rue T (1998) Rang des systèmes dynamiques gaussiens. Isr J Math 104:261–283
182. de la Rue T (1998) L'ergodicité induit un type spectral maximal équivalent à la mesure de Lebesgue. Ann Inst H Poincaré Probab Statist 34:249–263
183. de la Rue T (1998) L'induction ne donne pas toutes les mesures spectrales. Ergod Theor Dyn Syst 18:1447–1466
184. de la Rue T (2004) An extension which is relatively twofold mixing but not threefold mixing. Colloq Math 101:271–277
185. Ryzhikov VV (1991) Joinings of dynamical systems. Approximations and mixing. Uspekhi Mat Nauk 46 5(281):177–178 (in Russian); Math Surv 46(5)199–200
186. Ryzhikov VV (1992) Mixing, rank and minimal self-joining of actions with invariant measure. Math Sb 183:133–160, (Russian)
187. Ryzhikov VV (1994) The absence of mixing in special flows over rearrangements of segments. Math Zametki 55:146–149, (Russian)
188. Ryzhikov VV (1999) Transformations having homogeneous spectra. J Dyn Control Syst 5:145–148
189. Ryzhikov VV (2000) The Rokhlin problem on multiple mixing in the class of actions of positive local rank. Funkt. Anal Prilozhen 34:90–93, (Russian)
190. Ryzhikov VV, Thouvenot J-P (2006) Disjointness, divisibility, and quasi-simplicity of measure-preserving actions. Funkt Anal Prilozhen 40:85–89, (Russian)
191. Ryzhikov VV (2007) Weak limits of powers, simple spectrum symmetric products and mixing rank one constructions. Math Sb 198:137–159
192. Sato K-I (1999) Lévy Processes and infinitely divisible distributions. In: Cambridge Studies in Advanced Mathematics, vol 68. Cambridge University Press, Cambridge, pp 486
193. Schmidt K (1977) Cocycles of Ergodic Transformation Groups. In: Lecture Notes in Math. 1. Mac Millan, India
194. Schmidt K (2002) Dispersing cocycles and mixing flows under functions. Fund Math 173:191–199
195. Schmidt K, Walters P (1982) Mildly mixing actions of locally compact groups. Proc London Math Soc 45(3)506–518
196. Sinai YG (1994) Topics in ergodic theory. Princeton University Press, Princeton
197. Shklover M (1967) Classical dynamical systems on the torus with continuous spectrum. Izv Vys Ucebn Zaved Matematika 10:(65)113–124, (Russian)
198. Smorodinsky M, Thouvenot J-P (1979) Bernoulli factors that span a transformation. Isr J Math 32:39–43
199. Stepin AM (1986) Spectral properties of generic dynamical systems. Izv Akad Nauk SSSR Ser Mat 50:801–834, (Russian)
200. Tikhonov SV (2002) On a relation between the metric and spectral properties of $\mathbb{Z}^d$-actions. Fundam Prikl Mat 8:1179–1192
201. Thouvenot J-P (1995) Some properties and applications of joinings in ergodic theory. London Math Soc Lect Note Ser 205:207–235
202. Thouvenot J-P (2000) Les systèmes simples sont disjoints de ceux qui sont infiniment divisibles et plongeables dans un flot. Coll Math 84/85:481–483
203. Ulcigrai C (2007) Mixing of asymmetric logarithmic suspension flows over interval exchange transformations. Ergod Theor Dyn Syst 27:991–1035
204. Walters P (1982) An Introduction to Ergodic Theory. In: Graduate Texts in Mathematics, vol 79. Springer, Berlin, pp 250
205. Wysokińska M (2004) A class of real cocycles over an irrational rotation for which Rokhlin cocycle extensions have Lebesgue component in the spectrum. Topol Meth Nonlinear Anal 24:387–407
206. Wysokińska M (2007) Ergodic properties of skew products and analytic special flows over rotations. In: Ph D thesis. Toruń, 2007.
207. Veech WA (1978) Interval exchange transformations. J Anal Math 33:222–272
208. Veech WA (1982) A criterion for a process to be prime. Monatshefte Math 94:335–341
209. Veech WA (1982) Gauss measures for transformations on the space of interval exchange maps. Ann Math 115(2):201–242
210. Veech WA (1984) The metric theory of interval exchange transformations, I Generic spectral properties. Amer J Math 106:1331–1359
211. Vershik AM (1962) On (1962) the theory of normal dynamic systems. Math Sov Dokl 144:625–628
212. Vershik AM (1962) Spectral and metric isomorphism of some normal dynamical systems. Math Sov Dokl 144:693–696
213. Yassawi R (2003) Multiple mixing and local rank group actions. Ergod Theor Dyn Syst 23:1275–1304
214. Zeitz P (1993) The centralizer of a rank one flow. Isr J Math 84:129–145

Spin Dependent Exchange and Correlation in Two-Dimensional Electron Layers

M. W. CHANDRE DHARMA-WARDANA
Institute of Microstructural Sciences,
National Research Council of Canada, Ottawa, Canada

Article Outline

Glossary

Atomic units, a. u. The electron-charge $|e|$, and the mass m_e are taken as unity. The unit of time is fixed by setting the Plank constant $\hbar$ to unity. The Bohr radius $a_0 = \hbar^2/(m_e e^2)$ is one a.u. of length in the centimeter-gram-second(CGS) system which uses the "esu"(electrostatic unit of charge). The SI system uses the meter, kilogram, second (with the Ampere as the unit of current), $a_0 = (4\pi\varepsilon_0\hbar^2/(m_e e^2)$ where ε is the electric permittivity of the vacuum. The value of a_0 is 5.29177×10^{-9} cm. The unit of energy, the Hartree, is e^2/a_0 in the CGS system, being 27.2116 eV. In semiconductor physics, *effective atomic units* are used, with e^2/ε, replacing e^2, where ε is the dielectric constant. The band mass m_b is used instead of m_e, so that the effective Bohr radius $a_0^* = a_0(\varepsilon/m^*)$, where m^* is the effective mass. Then the effective Hartree is of the order of millivolts.

Confining potential This potential keeps the electron in a given spatial region. It is due to the physical structure of the device, the applied gate voltages etc.

Correlation energy The contribution to the total energy beyond the Hartree–Fock approximation, denoted by E_c.

Correlation hole is the depletion of electron density near an electron due to Coulomb repulsion effects.

Coupling constant is the ratio of the potential energy(PE) to the kinetic energy (KE). In a classical electron fluid, the KE (thermal energy) is T, and the PE is $1/r_s$ (atomic units), and r_s is the Wigner–Seitz radius. The coupling constant $=$ PE/KE $= \Gamma = 1/(r_s T)$. In quantum systems, the Fermi energy E_F is used instead of T for the KE, and $\Gamma = r_s$.

Classical-map hyper-netted-chain (CHNC) A method for using the classical hyper-netted-chain equation to calculate the correlation functions of quantum systems.

Effective mass is denoted by m^*, and is the band mass m_b in units of the electron mass m_e.

Exchange energy The part of the Hartree–Fock energy due to electron exchange, i. e., the "Fock" part, denoted by E_x. It is first order in the Coulomb interaction.

Fermi hole This denotes the reduction in the probability of finding a like-spin electron near another, due to Fermi statistics.

Hartree–Fock Hartree's self-consistent one-body approximation for interacting electrons is based on a product wavefunction. Fock included exchange using an antisymmetrized product. "Hartree–Fock" is the label for calculations of the energy, wavefunctions etc., where the electron moves in this mean potential generated by the electrostatics and the exchange effects. The Hartree term is zero in uniform systems.

HIGFET
Heterojunction-Insulated-Gate Field-effect Transistor.

Jellium A model "metal" where the positive ionic charges are replaced by a uniform static charge which neutralizes the free-electron charge.

MOSFET Metal-oxide semiconductor field-effect transistor.

Heterojunction A semiconductor-interface involving two dissimilar materials.

Hyper-netted-chain (HNC) A classical integral equation due to van Leeuwen, Groenveld and de Boer (1959) which non-perturbatively sums "hyper-netted-chain" diagrams, going beyond mean-field theory.

Electron gas parameter $\mathbf{r_s}$ See Wigner–Seitz radius.

Pair-correlation function Denoted by $h(\vec{r}) = g(\vec{r}) - 1$ where $g(r)$ is the pair-distribution function(PDF).

Pair-distribution function The pair-distribution function(PDF), $g(\vec{r})$, is the probability of finding a particle at the location $\vec{r}$, given a reference particle at the origin.

Plasma analogy A class of methods for approximately replacing a charged quantum fluid by an equivalent classical fluid at a finite temperature.

Pseudo-spin Discrete degrees of freedom beside the electron spin. The electrons in Si/SiO$_2$ interfaces occupy two valleys. The valley index is a pseudospin.

Quantum-Monte Carlo (QMC) In molecular dynamics (MD), Newton's equations of motion are integrated using a stochastic scheme based on the Metropolis algorithm. In QMC a trial wavefunction provides a probability measure for the Metropolis algorithm. The wavefunction is optimized in various ways, leading to "variational QMC", where the nodes of the trial wavefunction are held fixed. In "Diffusion QMC", the nodes are also relaxed.

Random-phase approximation (RPA) A time-dependent self-consistent field method where an electron with momentum $\vec{k}$ moves in an effective potential which contains the external potential and a $\vec{k}, \omega$ dependent screened potential. It reduces in the static $k \to 0$ limit to Thomas-Fermi screening, (or Debye–Hŭkel screening in classical systems). It is also called the "ring sum" or "bubble sum", and contains no exchange effects.

Subbands The electrons with the z-motion confined to a quantum well have discrete energy levels (index n). Each level carries with it a band of energies for the x, y motion. These are energy "subbands".

Singwi–Tosi–Land–Sjölander (STLS) A method due to Singwi et al. for determining the density-density correlation function of electrons (and other quantum systems) by truncating the equation of motion via an intuitive decoupling scheme involving the PDFs. STLS has been extended by Vashista, Ichimaru and others.

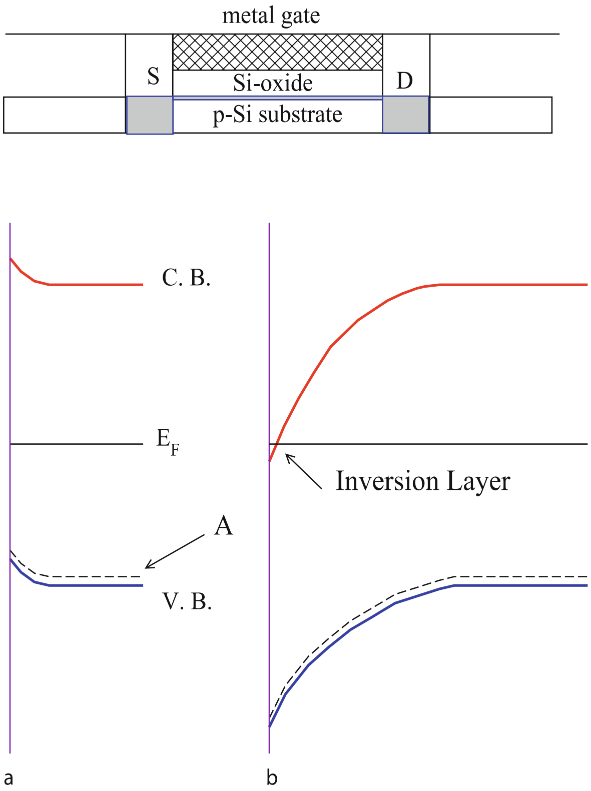
Wigner–Seitz radius In 2D this is the radius, denoted by r_s, of the circle containing, on the average, just one electron.

Definition of the Subject

Since the advent of density-functional theory (DFT), the exchange-correlation energy E_{xc} of an interacting system has become a basic quantity in many-particle theory. Here we study the E_{xc} of two-dimensional (2D) electron layers. Such layers contain electrons which move in the x and y directions, while confined in z. 2D layers are formed at insulator-semiconductor interfaces in heterojunctions, and more particularly at metal-oxide-semiconductor(MOS) interfaces. These include two types of semiconductors (e. g., GaAs and the alloyed form $Al_xGa_{1-x}As$, containing a small fraction x of Al, and written as AlGaAs for brevity). The interface region defines a "confining potential" where an electron layer may from [2]. SiO_2 is an insulator with a large bandgap, while Si can be doped in a controlled manner to behave as a conductor. The Si/SiO_2 interface supports the formation of an electron layer at the interface. The electron density n in such layers can be controlled using external potentials, and it is this which is the key to the importance of these materials systems. These structures are the basis of metal-oxide field-effect transistors (MOSFETS), ubiquitous in modern electronic devices (Fig. 1). The Si-MOSFET was developed in the 1960s, while it had been fore-shadowed as early as the 1930s in the work of Shockley and others.

Electron layers confined at the air-liquid interface of a fluid (e. g., Helium) were studied in the 1970s, but did not provide easy tunability of the density. High-mobility 2D systems in GaAs/GaAlAs heterostructures and in Si-MOSFETS ushered the more recent phase, unearthing new physics and new technologies, with the quantum Hall effect(QHE) [4], discovered in 1980, providing a classic example of fundamental physics directly leading to practical applications in metrology. Modern nanotechnology, spintronics, plasmonics etc., greatly dependent on the theory of electron layers.

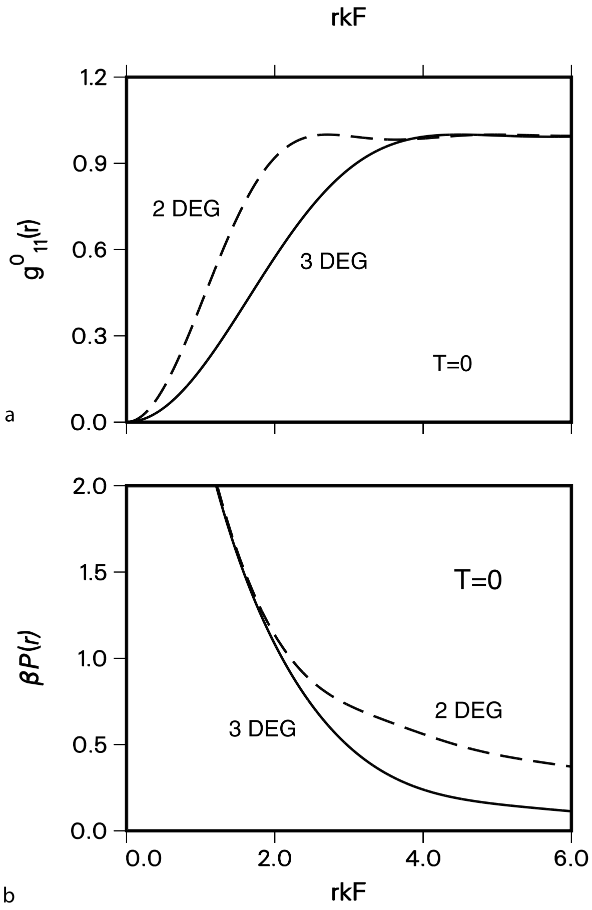
Graphite is made of 2D sheets of carbon atoms, i. e., "graphene" sheets, held together by weak inter-layer bonds The 2D electrons in graphene behave as massless Fermions with $\epsilon(\vec{k}) = \hbar v_F \vec{k}$, where v_F is the Fermi velocity. The hexagonal unit cell of graphene contains two inequivalent Carbon atoms. Thus there are two degenerate conduction bands and two valance bands, with a zero band gap located at two inequivalent points, labeled **K** and **K**′, in the hexagonal 2D-Brillouin zone of graphene. This system needs a spin index, a pseudospin index and also a band index. Graphene is a rich 2D system with novel physics (e.g, QHE) and the promise of new technological applications [9].



Spin Dependent Exchange and Correlation in Two-Dimensional Electron Layers, Figure 1

Top: A cross section of a *p*-type MOSFET. The *n*-type contacts labeled S (source) and D (drain) are made by ion-implantation into the *p*-type substrate. A voltage on the metal gate controls the current between S and D flowing in the Si/SiO_2 interface (channel region). Panel **a** shows the conduction band (c.b.), the acceptor levels (A), and the valance band (v.b.) modified by the application of negative bias to the metal gate. A sufficiently strong negative bias can produce an accumulation layer of holes. In **b** a sufficiently positive gate voltage is applied, bending the conduction band *below* the Fermi energy E_F, and creating a potential well which supports a 2D layer of electrons, known as an 'inversion layer'

The Coulomb interaction conspires to destroy the "single-particle" (i. e., "free" particle) picture of electron systems. Only the kinetic energy and the confining potential (called the "external potential") are needed in the simplest models. Mean-field (e.g, Hartree–Fock) models provide an "effective" single-particle picture. The "many-body effects", which transcend the simple picture, lead to new effects like plasmons, charge density waves, spin-polarized states, Wigner crystallization, fractional-QHE, and super-



Spin Dependent Exchange and Correlation in Two-Dimensional Electron Layers, Figure 2

***Upper panel* shows the pair-distribution function (PDF) for non-interacting parallel-spin electrons in 2D and 3D electron gases. The depletion near the origin is the "Fermi hole". The *lower panel* shows the equivalent *classical* potential (Pauli exclusion potential, in $\beta = 1/T$ units) which would create such a depletion in the PDF at the temperature T (at $T = 0$, only the product $\beta P(r)$ is meaningful). The x-axis is the electron-pair separation in reciprocal k_F units**

conductivity [18]. Many-body effects are more enhanced in low dimensional (e. g., 1D and 2D systems in comparison to 3D) systems. The Coulomb interactions force two electrons to repel each other and form a "Coulomb hole" or correlation hole, i. e., a region of low probability of approach. The energy consequence of this is the "correlation energy". Similarly, since the Pauli principle forbids the presence of two non-interacting electrons of the same spin in a common spatial eigenfunction, a "Fermi hole " is formed and leads to an exchange energy (see Fig. 2).

There is no Fermi hole for antiparallel spins, and their non-interacting PDF, i. e., $g^0_{12}(r)$ is unity for all r. When interactions are included, a Coulomb hole is formed. Hence the evaluation of E_{xc} is directly related to obtaining good PDFs.

Thus the theory of E_{xc} is of great importance in modern density-functional many-body approaches. From a technological point of view, the subject is of critical importance in the proper design of modern nano-structure devices and quantum-well lasers [16].

Introduction

We use "effective" atomic units, with the effective Bohr radius a_0^* as the unit of length. The dielectric constant ε is of the order of 10–12 for common semiconductors, while m^* may be 0.06 in GaAs and 0.19 in Si. Hence the effective Hartree is measured in meV, unlike the atomic Hartree which is 27.12 eV. Thus the properties of the material enter into the theory only via ε, m^*, and the effective Landé factor g^*, and these are usually absorbed into the effective atomic units. Since the electrons reside in the interface between two materials (say A, B),the dielectric constant ε relevant to the electron layer has to be evaluated from the individual dielectric constants of A and B. This is trivial in the case of GaAs/AlGaAs systems, since the dielectric constants are very close. This is no longer the case with Si/SiO$_2$ systems where the Si and the Si-oxide have nominal dielectric constants of 11.5 and 3.9 respectively. Then an average dielectric constant has to be used, as discussed in the appendix of [2], and in [8].

If the z-motion of an electron is restricted to a finite length a_z, as in an inversion layer, the allowed quantum states form *discrete subbands*, with energies $E(n, k_x, k_y)$:

$$E_n(k_x, k_y) = \epsilon_n + \epsilon(k_x, k_y)\,.$$

The energy of the in-plane motion, $\epsilon(k_x, k_y)$, depends on the in-plane momenta $\hbar k_x$ and $\hbar k_y$ with $\vec{k} = \vec{k}_x + \vec{k}_y$. For materials like GaAs/AlGaAs, or Si/SiO$_2$ interfaces, the in-plane energy dispersion is essentially parabolic, with

$$\epsilon(k_x, k_y) = \frac{(\hbar k_x)^2 + (\hbar k_y)^2}{2m^*}\,.$$

Here m^* is an effective mass, assumed to be the same in the x and y directions Only the lowest subband, say $n = 1$, is occupied in a 2D layer. Let ϵ_1 be the energy zero. The highest occupied state is given by the in-plane momentum $\hbar k_{\text{F}}$, known as the *Fermi momentum*. The Fermi energy (ϵ_{k_F}) has to be far below the bottom of the second subband to ensure a 2D layer. The electron density, i. e., the number of electrons n per unit area has n_1 "up-spin", and n_2 "down-spin" electrons. If $n_1 \neq n_2$ we have a *spin-polarized* electron system. The *degree of spin polarization*

ζ is:

$$\zeta = \frac{n_1 - n_2}{n_1 + n_2} \,.$$

Thus the state of a clean 2D-electron layer may be specified by n, ζ, and the temperature T. Additional pseudo-spin or valley indices are sometimes necessary.

The Fermi Energy and the Role of r_s as a Coupling Constant

When the electron-number density per a.u. of area is n, the Wigner–Seitz radius r_s is given by $r_s = 1/\sqrt{(\pi n)}$. We also have $n_1 = nx_1$, $x_1 = (1+\zeta)/2$, $n_2 = nx_2$, $x_2 = (1-\zeta)/2$. The Fermi momentum $k_{F\sigma}$ of a given spin species σ is evaluated by requiring that the sum of all occupied states of the species σ adds upto the density n_σ.

$$\int_0^{k_F} 2\pi \frac{k\mathrm{d}k}{(2\pi)^2} = n_\sigma \,.$$

This gives $k_\sigma = (4\pi n_\sigma)^{1/2}$. Since $n = n_1 + n_2$, the overall $k_F = (2\pi n)^{1/2}$, i. e., $k_F = \sqrt{2}/r_s$ and $E_F = k_F^2/2 = 1/r_s^2$.

Hence, the "coupling strength", i. e., ratio of the Coulomb energy $(1/r_s)$ to the kinetic energy $(\sim E_F)$ at $T = 0$ is clearly r_s. High density systems ($r_s < 1$) are weakly coupled, while low density systems (high r_s) are strongly correlated. When $r_s \sim 26-27$, the 2D electron system acquires a ferromagnetic ground state. At stronger coupling, $r_s \simeq 35$, a 2D Wigner crystal is formed. The strength of the 2D massless electron interactions in graphene cannot be specified by an r_s since a Bohr radius cannot be specified. Instead, the coupling constant g is taken as the ratio of a typical Coulomb energy $(1/a_0)$ to the hopping energy (t) on the Hexagonal 2D unit cell with a lattice constant of a_0.

$$g = \frac{e^2/a_0\varepsilon_0}{\hbar E_F} = \left\{e^2/a_0\varepsilon_0\right\}\left\{t\sqrt{3/2}\right\} \,.$$

If $\varepsilon_0 = 1$, then $g \simeq 2.7$. Thus, when the graphene sheet is placed on most substrates, $\varepsilon > 1$, and the electron interactions become weak, i. e., $g \leq 1$.

Fourier Transforms

The Fourier transform of 2D-functions relating their r-space forms with q-space forms is very useful. If the 2D electron system is uniform in all directions of the $x-y$ plane, i. e., isotropic, then we only need *radial* Fourier transforms. These are obtained using the relations:

$$F(\vec{r}) = \int_0^\infty F(\vec{q}) J_0(qr) q\mathrm{d}q/(2\pi)$$

$$F(\vec{q}) = \int_0^\infty F(\vec{r}) J_0(qr) 2\pi r\mathrm{d}r \,.$$

Here $J_0(qr)$ is a Bessel function [1] such that:

$$J_0(qr) = 2\int_0^\pi cos(qr\cos\theta)d\theta \,.$$

Then it is easy to show that the Coulomb interaction $1/r$ has the Fourier form $2\pi/q$. Here we assume ideally thin layers (unlike in Sect. "Graphene: 2D Two-Valley System on a Honey-Comb Lattice"). A screened Coulomb interaction $V_s(r)$, with an exponential damping is often encountered. This potential has an analytic Fourier transform and is:

$$V_s(r) = \exp(-k_s r)/r \qquad V_s(q) = 2\pi/\left\{\left(q^2 + k_s^2\right)^{1/2}\right\} \,.$$

The potential contains a "screening wavevector" k_s, and appears in what is known as "Thomas-Fermi" screening. This type of potential is also called a "Yukawa potential". The 'range' of the potential is of the order of $1/k_s$, and is known as the screening length.

The Hamiltonian of the System

The total Hamiltonian contains a contribution from a *uniform* neutralizing background of positive charge n_b which is static. In a real system, e. g., a metal, the neutralizing background is provided by the ion subsystem. Each ion has a short-ranged core-region and a Coulomb-like long-range potential which overlaps the long-range potential of the neighbours. In many metals (e. g., Sodium) the resulting overlapping positive potential is, to a very good approximation, similar to a smudged-out, structureless "jellium". In a plasma, the positive ions are dynamical but extremely heavy, and hence a jellium approximation, known as the one-component plasma (OCP) is often used. The jellium model is used in theories of electron layers as the physics then focuses entirely on electron-electron effects.

We write the electron density operator and the Coulomb interaction as:

$$n(\vec{r}) = \sum_i \delta(\vec{r} - \vec{r}_i) \,, \qquad V_{ee} = \frac{1}{2}\sum_{ij} \frac{n(\vec{r}_i)n(\vec{r}_j)}{|\vec{r}_i - \vec{r}_j|} \,.$$

The latter has a divergence when $i \to j$, with the electron interacting with "itself". We also have the terms V_{bb} and V_{be}, involving the background. For example,

$$V_{\mathrm{be}} = -\sum_{ij} n_{\mathrm{b}} \frac{n(\vec{r}_j)}{|\vec{r}_i - \vec{r}_j|}$$

where $n_{\mathrm{b}} = n$ is the uniform-background density. The background term V_{bb} has a divergent self-interaction term similar to that in V_{ee}, with (positive) sign. However, the interaction of the electron charge with the background, i. e., V_{be}, has a negative divergent term, and cancels both divergent (positive) self-interaction terms. Thus the total Hamiltonian, inclusive of the background is free of divergences. In Fourier space, the coulomb interaction is $2\pi/q$, and hence the divergence manifests in the limit $q \to 0$. This is exactly canceled by the corresponding $q \to 0$ terms in the background. Also, since the background is uniform, there are no other Fourier components. The above discussion involves the "subtraction" of divergent quantities, and can be made more rigorous using the screened Coulomb potential $V_{\mathrm{s}}(r)$, and taking the limit $k_{\mathrm{s}} \to 0$ at the end.

The electron number-density operator $n(r)$ in real space has a Fourier transform $n(q)$, defining the occupation number in the momentum state q. The expectation value of $n(r)$ is of course the constant n, which may be written as $\overline{n}$ where needed.

The momentum eigenstates of the uniform noninteracting system are the complete set of plane waves.

$$\phi_q = \phi(\vec{q}) = \mathrm{e}^{\vec{q}\cdot\vec{r}}/\Omega^{1/2}$$

Here we use a normalization volume Ω, which is sometimes set to unity, or the infinite-volume limit is taken, when the normalization becomes $1/(2\pi)^{1/2}$. The occupation number in each momentum state, for non-interacting Fermions at a temperature T, and chemical potential μ is given by:

$$n_k = \frac{1}{1 + \mathrm{e}^{\beta(\epsilon_k - \mu)}} \, .$$

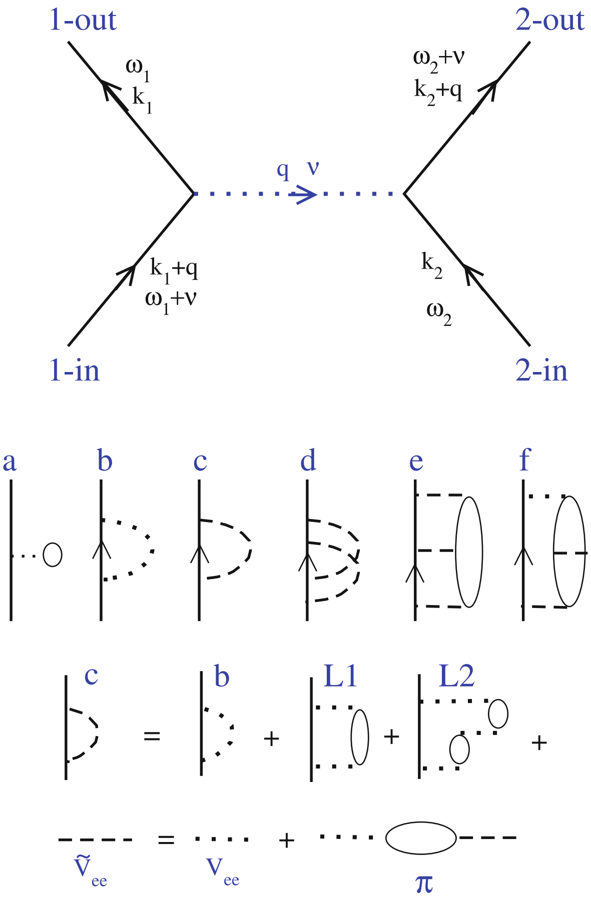
Here $\beta = 1/T$ where T is expressed in energy units. At $T = 0$, the chemical potential μ becomes the Fermi energy E_{F}, and the occupation numbers reduce to unity for energies $\epsilon_k < E_F$, and zero for higher energies. The occupation number n_k given above is really the expectation value of the operator n_k. When we need to emphasize this distinction, we write the operator as $\hat{n}_k$, and the meanvalue as $\overline{n}_k$. In a uniform system, most properties are dependent only on the modulus $|k|$ of the momentum $\vec{k}$, and we sometimes suppress the vector notation for brevity. We also introduce creation and annihilation operators a_k^+ and a_k which add an electron to a momentum state $\vec{k}$, or remove an electron from the momentum state $\vec{k}$. Thus the number operator acting on the eigenstate $|\phi_k\rangle$ is given as:

$$\hat{n}_k|\phi_k >= a_k^+ a_k|\phi_k\rangle = \overline{n}_k|\phi_k\rangle \, .$$

The total Hamiltonian of the system is:

$$H = (H^0 - \mu N) + V'_{\mathrm{ee}} \, .$$

Here we are using a thermal ensemble which allows for the definition of a chemical potential μ and a temperature T. Also, $N = \sum_{\vec{k}} \hat{n}_k$ is an operator fixing the mean



Spin Dependent Exchange and Correlation in Two-Dimensional Electron Layers, Figure 3
***Top Panel*: Feynman diagram for the bare Coulomb interaction $V_{\mathrm{ee}}(q)$ of two electrons. *Bottom panel*: a shows the exchange interaction which is first order in the bare interaction $V_{\mathrm{ee}}(q)$. b is the bare exchange, and c has the RPA "bubble sum", where $\tilde{V}_{\mathrm{ee}}(q,\nu)$ is given by an integral equation (a geometric series in this case) involving the polarization "bubble" $\pi(q,\nu)$. Contributions from diagrams like c–d are not included in the RPA propagators**

number of particles $\langle N\rangle$ in the volume Ω, given μ and T. The Coulomb interaction V'_{ee} carries a prime to indicate that a jellium background has been included to remove the divergencies at zero momentum transfer. Then we have, in momentum space:

$$H = \sum_{\vec{k}} (\epsilon_k - \mu) a_k^+ a_k + \frac{1}{2\Omega} \sum_{\vec{k}_1 \vec{k}_2 \vec{q}}{}' V_q a_{k_1}^+ a_{k_2+q}^+ a_{k_2} a_{k_1+q} .$$

The prime on the summation indicates that the case $q = 0$ is excluded, as it is exactly cancelled by the jellium background. The suffixes on the cluster of four operators are such that the total momentum of the two electrons is conserved during the interaction. That is, an electron in the initial state $\vec{k}_1 + \vec{q}$ is in the final state $\vec{k}_1$, while the second electron, in the initial state $\vec{k}_2$ is transfered to the state $\vec{k}_2 + \vec{q}$. The momentum transfer $\vec{q}$ is effected through the Coulomb interaction $V_q = 2\pi/q$. The momentum is conserved in the Coulomb collision since we are dealing with a *uniform system* (see top panel in Fig. 3).

Ideally Thin 2D Electron Layers

This system is most closely approximated in Si/SiO$_2$ systems, or in specially fabricated GaAs/GaAlAs quantum wells (HIGFETs tend to yield thick 2D electron layers and are discussed in Sect. "2D Layers with Finite Thickness"). Microscopic many-body theories have used this system as the basic "work horse". Early work used diagrammatic and other perturbation methods, or truncated equation of motion methods, like STLS, developed by Singwi and coworkers. The objective of the latter is to develop PDFs non-perturbatively, and calculate the exchange and correlation energies from an integration over the coupling constant (described below). However, the perturbation methods (restricted to the high-density regime in validity) provide basic reference results.

A Brief Outline of the Diagrammatic Method

Much of the initial work was based on a perturbation expansion of the energy or the electron propagator in powers of the Coulomb interaction $V_{\mathrm{ee}}(q)$, as in Fig. 3. In the graphs labeled a–f we look at the self-energy (i.e, the effective potential) of an electron propagating through the interacting fluid. In (a) an electron propagates forward, while the interaction with the second electron occurs with zero-momentum transfer ($q = 0$), i.e, one may imagine that the "2-out" line "falls into" the "2-in" line to from a bubble, when $\vec{k}_2 + \vec{q} = \vec{k}_2$ by momentum conservation. Such $q = 0$ terms are Hartree terms, and have already been eliminated via the static background. The graph (b) is the first-order exchange diagram (the "Fock" term). This may be pictured as the "1-out" line "falling" into the "2-in" line, exchanging electron identities. This is possible only if both electrons have the same spin. The graph (c) is the *screened exchange* diagram where the Coulomb interaction is *dynamically screened* by summing the simple polarization loops L1, L2 etc. This is known as the RPA-screened exchange, where RPA (random-phase approximation) is a name originating from methods of derivation using self-consistent equations. However, each term in such an expansion is infinite; e. g., the second-order perturbation term L1, and those beyond it (L2 etc.) are all divergent, as they involve integrals of the form $\int (2\pi/q)^n q \mathrm{d}q \ldots$, where $n \geq 2$ (there are standard "Feynman rules" for converting the graphs shown in Fig. 3 into algebraic expressions [18]; they are not needed for our purpose). The divergence arises from perturbation terms containing simple polarization loops (denoted by $\pi(q, \omega)$ in Fig. 3) connected by two or more direct interaction lines. However, although each term is infinite, the *sum* L1 + L2 + ... is finite. The bare interaction $V_{\mathrm{ee}}(q, \omega)$ gets "screened" by the polarization processes denoted by the bubble (particle- hole pairs), and gives the well-behaved (non-divergent) $\tilde{V}_{\mathrm{ee}}(q)$ which is used in the graph (c). The contributions from scattering processes shown in Fig. 3d–f, etc., are *not* included in the RPA sum.

Thus the contribution to the effective potential seen by an electron at the Fermi level can be written as

$$V_{\mathrm{eff}} = V_{\mathrm{xc}} = V_{\mathrm{x}} + V_{\mathrm{c}}$$

where V_{x} is the exchange potential, arising from the first-order diagram Fig. 3b, while V_{c} is the contribution from *all other* higher processes. Thus V_{c} in the RPA is simply the sum of contributions from L1, L2 etc. A better approximation would be a V_{c} which involves Fig. 3e–f, and all other terms. But the numerous higher terms are virtually impossible to evaluate, and their partial inclusion is problematic as various sum rules, Ward identities etc., have to be satisfied [18]. In analogy with the effective potential, we can also divide the total energy of the system (beyond the kinetic energy) as:

$$E_{\mathrm{xc}} = E_{\mathrm{x}} + E_{\mathrm{c}}; \quad \epsilon_{\mathrm{x}} = E_{\mathrm{x}}/N; \quad \epsilon_{\mathrm{c}} = E_{\mathrm{c}}/N$$

where E_{x} is the exchange energy, and E_{c} is the correlation energy. The latter by definition contains all contributions other than exchange. A large part of the spin dependence of E_{xc} is contained in E_{x}. The RPA is an evaluation of E_{c} via the series L1, L2 etc.

The Adiabatic Connection Formula

Another approach to E_{xc} is to begin with the non-interacting system and "build-up" the interacting system by increasing the value of the "coupling constant" λ from $0 \to 1$ in the scaled Coulomb potential

$$V_{\mathrm{ee}}(\lambda, r) = \lambda / r\,.$$

The usual coupling parameter r_{s} is here assumed to be merely a density parameter ($r_{\mathrm{s}} = 1/\sqrt{\pi n}$) held constant during this "charging process". As λ increases to unity, the PDFs $g_{ij}(\lambda, r)$, where i, j indicate spin species, change from the non-interacting $g^0_{ij}(r)$, Fig. 2, to the interacting from $g_{ij}(r)$. This is accompanied by the "build up" of the energies arising from the interactions. Thus the exchange-correlation energy E_{xc} appears as an integration over the coupling constant (if we are dealing with a system at a finite temperature T, then we have Helmholtz free energies F_{xc} instead of E_{xc}). Thus, if x_1, x_2 are the fractional compositions n_i/n, then the exchange-correlation energy per particle E_{xc}/N is

$$E_{\mathrm{xc}}/N = \int_0^1 \frac{\mathrm{d}\lambda}{\lambda}\frac{n}{2} \int 2\pi r \mathrm{d}r \frac{\lambda}{r} \sum_{ij} x_i x_j (g_{ij}(\lambda, r) - 1)\,. \quad (1)$$

This is known as the adiabatic connection formula [11], and is due to Pauli. Thus if the fully interacting PDF is known, the exchange-correlation energy can be obtained without the limitations of perturbation theory. The full PDF can be evaluated by (i) Quantum Monte Carlo methods (QMC), (ii) integral-equation methods developed from correlated wavefunctions using Feenberg-type analyses, (iii) classical-map hyper-netted chain (CHNC) methods, (iv) methods for specific applications as in Giorgi et al. [10] (v) SLTS methods for moderately coupled fluids.

The Exchange Energy

Note that the exchange energy alone is given by

$$E_{\mathrm{x}}/N = \frac{n}{2} \int 2\pi r \mathrm{d}r v(r) \sum_{ij} x_i x_j \left(g^0_{ij}(r) - 1\right)\,. \quad (2)$$

That is, the Fermi hole of the non-interacting PDF determines the exchange energy. This was *defined* as a contribution taken to *first order*, Fig. 3b. Since one factor of $v(r)$ is already included in Eq. (2), we can only use $g^0_{ij}(r)$ as a better form would have go beyond first order in V_{ee}. The noninteracting PDF $g^0_{ij}(r)$ can be evaluated [18] directly from the noninteracting one-body wavefunctions (plane waves):

$$\begin{aligned} h_{ij} &= g_{ij} - 1 \\ h^0_{ij}(r) &= -\delta_{ij} F_i(r) F_j(r) \\ F_i(r) &= \frac{1}{n_i} \int \frac{\mathrm{d}\vec{k}}{(2\pi)^2} f_i(k) \mathrm{e}^{i\vec{k}\cdot\vec{r}}\,. \end{aligned}$$

Here $f_i(k)$ is the Fermi occupation number for the momentum state k. At $T = 0$, $F_i(r) = 2J_1(k_i r)/(k_i r)$; $k_i = \sqrt{4\pi n_i}$ where $J_1(x)$ is a Bessel function [1], and k_i is the Fermi momentum of the spin species i. At finite temperatures, the integrations have to be done numerically.

Formulae for the Kinetic and Exchange Energies.

At zero temperature, the Hartree–Fock energy can be written in terms of the spin polarization ζ and the density parameter r_{s}. The Hartree energy E_{H} is zero.

$$E_{\mathrm{HF}} = E_0 + E_{\mathrm{H}} + E_X\,; \quad E_{\mathrm{H}} = 0 E_0/N = \frac{1+\zeta}{2r_{\mathrm{s}}^2}\,;$$

$$E_{\mathrm{x}}/N = -\frac{2\sqrt{2}}{3\pi r_{\mathrm{s}}}\left[(1+\zeta)^{3/2} + (1-\zeta)^{3/2}\right]$$

The spin-dependent energies, e. g. $E_{\mathrm{x}}(\zeta)$, may be written as:

$$\begin{aligned} E_{\mathrm{x}}(\zeta) &= E_{\mathrm{x}}(0) + (E_{\mathrm{x}}(1) - E_{\mathrm{x}}(0))P(\zeta) \\ P(\zeta) &= \frac{\zeta_+^\alpha + \zeta_-^\alpha - 2}{2^\alpha - 2}\,. \end{aligned}$$

Here $\zeta_\pm = (1 \pm \zeta)$, and $P(\zeta)$ is called the "polarization function"; $\alpha = 3/2$ in Hartree–Fock theory.

The Correlation Energy

The calculation of the correlation energy E_{c} is the heart of the many-body problem, and is the main challenge. If E_{c} were known, the total ground state energy is $E_{\mathrm{T}} = E_{\mathrm{HF}} + E_{\mathrm{c}}$. A widely used QMC evaluation was given in 1989 by Tanatar and Ceperley [23] using a variational Monte Carlo method (VMC) as well as a Greens Function Monte Carlo method. They gave parametrized forms for the unpolarized ($\zeta = 0$) and fully polarized ($\zeta = 1$) energies per electron, i. e., $\epsilon_{\mathrm{c}}(r_{\mathrm{s}})$, in the form

$$\epsilon_{\mathrm{c}}(r_{\mathrm{s}}) = a_0 \frac{1 + a_1 x}{1 + a_1 x + a_2 x^2 + a_3 x^3}$$

where $x = \sqrt{r_{\mathrm{s}}}$. The parameters have the values, $a_0 \ldots a_3 = -0.1784,\ 1.1300,\ 0.9052,\ 0.4165$ for $\zeta = 0$ and $-0.02575,\ 340.5813,\ 75.2293,\ 37.0170$ for $\zeta = 1$. No simulations were done for intermediate ζ, and hence it was often *assumed* that the exchange-like polarization function $P(\zeta)$ could be used for the correlation energy as well.

$$\epsilon_{\mathrm{c}}(\zeta) = \epsilon_{\mathrm{c}}(0) + (\epsilon_{\mathrm{c}}(1) - \epsilon_{\mathrm{c}}(0))P(\zeta)$$

A more detailed study, given by Attaccalite et al. in 2002 is currently the best available QMC form of E_{c}. These authors [3] did simulations for finite-ζ as well and proposed a parameterized form.

$$\begin{aligned}
\epsilon_{\mathrm{c}}(r_{\mathrm{s}},\zeta) &= (\mathrm{e}^{-\beta r_{\mathrm{s}}} - 1)\epsilon_x^6(r_{\mathrm{s}},\zeta)\\
&= +\epsilon_{\mathrm{c}}(r_{\mathrm{s}},0) + \alpha_1(r_{\mathrm{s}})\zeta^2 + \alpha_2\zeta^4\\
\alpha_1(r_{\mathrm{s}}) &= (1/2\left[\partial^2\epsilon_{\mathrm{c}}(r_{\mathrm{s}},\zeta)/\partial^2\zeta^2\right]_{\zeta=0}\\
\alpha_2(r_{\mathrm{s}}) &= (1/24)[\partial^4\zeta^4]_{\zeta=0}\\
\epsilon_x^6(r_{\mathrm{s}},\zeta) &= \epsilon_x(r_{\mathrm{s}},\zeta) - \{1 + (3\zeta^2)(1 + \zeta^2/16)\}
\end{aligned}$$

Here parameters α_1, α_2 are spin-stiffness constants. A short computer program for the calculation of ϵ_{c} is available [24].

CHNC Approaches

QMC procedures are computationally very demanding, and hence the development of other, simpler methods is of great interest. The accuracy of such methods can be ascertained by comparison with QMC results, and then such methods can be used in areas where QMC is prohibitive (e. g., at finite temperatures, many valley systems). In this respect, the classical-map HNC approach (CHNC) is worthy of note. In CHNC [21], the quantum fluid at $T = 0$ is replaced by a classical Coulomb fluid at a finite temperature T_{q}. The latter is chosen requiring that (i) the non-interacting classical fluid has the same $g^0_{ii}(r)$ as the quantum fluid. (ii) the correlation energy of the interacting classical fluid has the same correlation energy as the quantum fluid. The first requirement is easily met, and involves the construction of a classical potential which would give the same Fermi hole as the quantum fluid (see Fig. 2). The classical potential, $\beta V_{\mathrm{pau}}(r)$ is called the "Pauli potential" and is a universal function of rk_{F}. This potential is zero for anti-parallel spins. This procedure fixes a 'Pauli potential' but not the temperature T_{q}, since only the product $\beta \times V_{\mathrm{pau}}(r)$ is obtained. The second condition is used to determines T_{q} by evaluating $\epsilon_{\mathrm{c}}(r_{\mathrm{s}})$ using the adiabatic connection formula, for a trial value of T_{q}., and a chosen ζ, say $\zeta = 1$. Then T_{q} is adjusted till the QMC value of $\epsilon_{\mathrm{c}}(r_{\mathrm{s}}, \zeta = 1)$ is obtained. Thus a table of T_{q} for each r_{s}, i. e., $T_{\mathrm{q}}(r_{\mathrm{s}})$, is obtained. The classical PDFs needed here are easily calculated using a well established hyper-netted-chain technique. Only the ϵ_{c} value at $\zeta = 1$ is used as the input to the fit. The output is the full tabulation of the interacting $g_{ij}(r)$. It is found to be in *remarkably good agreement with the QMC generated* $g_{ij}(r)$. This is perhaps a consequence of DFT where it is asserted that the true density distribution is obtained if the true ground state energy (i.e, $E_0 + E_X + E_{\mathrm{c}}$) is captured. It is also found that this CHNC procedure yields accurate $\epsilon_{\mathrm{c}}(r_{\mathrm{s}})$ values at spin-polarizations different to the input value. The $T_q(r_{\mathrm{s}})$ function is found to be transferable to many Coulomb-fluid problems including hydrogen plasmas [6].

Fermi-Liquid Parameters

A knowledge of the correlation energy as a function of the r_{s}, ζ, and T enables an easy evaluation of Landau Fermi liquid parameters (LFLP). These are usually evaluated from complicated perturbation theory calculations whose domain of validity is, strictly speaking, restricted to small r_{s}. Three quantities are of great interest for the LFLP. These are, the inverse compressibility κ which is the density derivative of the chemical potential μ, the enhanced spin-susceptibility χ compared to the Pauli susceptibility χ_{P}, (incorporated in g^*, the effective Landé g factor), and the effective mass m^*. The latter involves the interacting and ideal specific heats C_v, C_v^0. AT finite T, the correlation energy is replaced by the correlation Free energy, $F_{\mathrm{c}}(r_{\mathrm{s}}, \zeta, T)$. This is the correction to the Helmholtz free energy beyond the Hartree–Fock approximation. If $F_{\mathrm{c}}(r_{\mathrm{s}}, \zeta, T)$ is known, a non-perturbative result becomes available for κ, m^* and g^* (for detailed, see [8]). Thus:

$$\begin{aligned}
m^* &= C_{\mathrm{v}}/C^0_{\mathrm{v}} = 1 + \frac{\left[\partial^2 F_{\mathrm{xc}}(t)/\partial t^2\right]}{\left[\partial^2 F_0(t)/\partial t^2\right]}\\
\chi_P/\chi_{\mathrm{s}} &= (m^* g^*)^{-1} = 1 + \frac{\left[\partial^2 F_{\mathrm{xc}}(\zeta)/\partial\zeta^2\right]}{\left[\partial^2 F_0(\zeta)/\partial\zeta^2\right]}\\
1/\kappa &= n^2\partial\mu/\partial n = n^2\partial^2[F_0 + F_{\mathrm{x}} + F_{\mathrm{c}}]/\partial n^2\,.
\end{aligned}$$

Calculations of the ground-state energy as a function of r_{s}, ζ, as well as the spin-susceptibility enhancement using the QMC and the CHNC methods show that the 2D fluid (which is paramagnetic $\zeta = 0$ at high densities), becomes a ferromagnetic liquid (see [3,7]) for $r_{\mathrm{s}} \simeq 26$. Thus the gain in kinetic energy due to spin-polarization is compensated by the exchange-correlation energy at sufficiently low densities. If simple perturbation methods are used for evaluating $E_{\mathrm{xc}}(r_{\mathrm{s}})$, incorrect predictions (e. g., a magnetic transition at $r_{\mathrm{s}} \sim 2$ in RPA) are obtained. The spin-susceptibility enhancement "blows up" towards infinity, as the transition is approached. These theoretical predictions have been confirmed experimentally (see [22]).

2D Layers with Finite Thickness

The assumption that the 2D layer is infinitely thin (i. e., “ideal”) is often incorrect in practice. The lowest subband of a quantum well supporting a 2D layer may have an z-extension w of 10–20 nm, with a density distribution $n(z)$. If the next subband is inaccessible to the electrons, then they are 2D systems with a finite thickness. The HIGFETS used in many experiments contain a nearly triangular potential well (see Fig. 1) and the lowest subband wavefunction is approximated by the Fang-Howard(FH) function [2] $\phi_{\mathrm{fh}}(z)$. The difficulty of modeling exchange and correlation in such systems using either the ideal-2D parametrized E_{xc} forms, or the 3D forms (to allow for the z-extension) was noted by Martin et al. [14]. In fact, new exchange-correlational functionals have to be constructed for the effective-2D potential found in thick-2D layers [5]. We denote the Coulomb potential in an infinitely thin layer by $V(r) = 1/r$, while $W(r)$ is used for the effective 2-D potential of a thick layer. The potential $W(r)$ between two electrons having coordinates $(\vec{r}_1, z_1)$ and $(\vec{r}_2, z_2)$, with $\vec{r} = \vec{r}_1 - \vec{r}_2$ is given by,

$$W(r) = \int_0^{z_m}\int_0^{z_m} \frac{\mathrm{d}z_1 \mathrm{d}z_2 n(z_1) n(z_2)}{\left[r^2 + (z_1 - z_2)^2\right]^{1/2}} . \tag{3}$$

Here z_{m} is ∞ for FH, while $z_{\mathrm{m}} = w$ for a quantum well. The potential $W(r) = (1/r)F(r)$ and the form factor $F(r)$ reflects the effect of the z-extension of the density. It has been shown that any arbitrary density distribution $n(z)$ can be replaced by a electrostatically equivalent rectangular slab of width w by choosing w such that: in the 2D plane as $n(z)$.

$$n_{\mathrm{cd}} = 1/w = \int n(z)^2 \mathrm{d}z . \tag{4}$$

The quasi-2D potential for a constant-density slab of width w is given by

$$W(r) = V(r)F(s), \quad s = r/w, \quad t = \sqrt{\left(1 + s^2\right)} \tag{5}$$

$$F(s) = 2s\left[\log\frac{1-t}{s} + 1 - t\right] . \tag{6}$$

This potential tends to $1/r$ for large r, and behaves as

$$\frac{2}{w}\left(\ln\frac{2w}{r} + \frac{r}{w} - 1\right)$$

for $r < w$. Thus the short-range behaviour is logarithmic and weaker than the Coulomb potential. The k-space form of the CDM potential is:

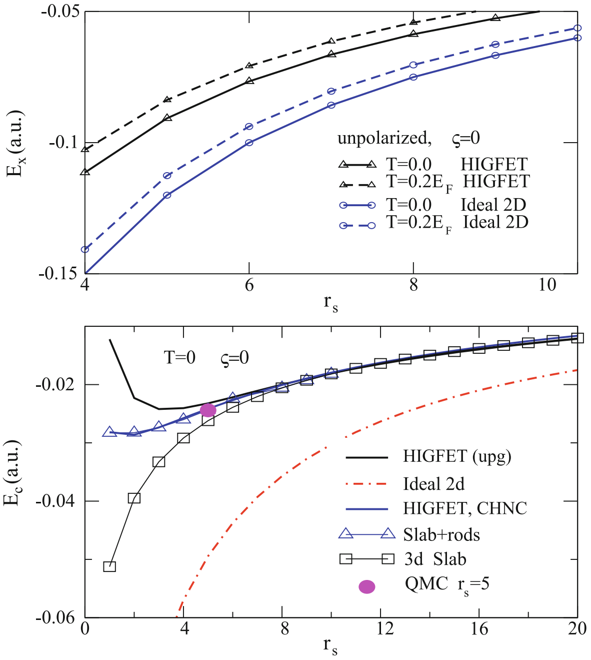
$$W(k, w) = V(k)F(p), \quad p = kw \tag{7}$$

$$F(p) = (2/p)\{(\mathrm{e}^{-p} - 1)/p + 1\} . \tag{8}$$

The form factors $F(s)$ and $F(p)$ tend to unity as $w \to 0$. The effective w in HIGFETS depends on r_{s}, i. e., in atomic units.

$$w = 16/(3b)b^3 = 33/\left(2r_{\mathrm{s}}^2\right) .$$

Using these potentials, layer-thickness dependent exchange-correlations functionals $E_{\mathrm{x}}c(r_{\mathrm{s}}, \zeta, w)$ can be evaluated using perturbation methods, or using the CHNC. The resulting xc-functionals have been given in [5], and displayed in Fig. 4. It is found that the exchange energy is reduced due to layer thickness which enters as w/r_{s}, thus



Spin Dependent Exchange and Correlation in Two-Dimensional Electron Layers, Figure 4

a The exchange energy E_{x} (Hartree a.u.) of a 2DES in a HIGFET compared to that of an ideal 2DES, at $T/E_{\mathrm{F}=0}$ and 0.2, with $\zeta = 0$. *Solid line with circles*, ideal 2D, $T = 0$, *Solid line with triangles*, HIGFET at $T = 0$. Corresponding *broken lines* are for $T = 0.2E_F$. b The correlation energy E_{c} at $T = 0$ and $\zeta = 0$, for the HIGFET layer. HIGFET(upg), *black solid line*, is the “unperturbed-g” approximation. The *deep grey dashed line*, HIGFET(CHNC), is the full calculation. This is compared with the correlation energy of a 3D slab model (*line with squares*), and the “slab+rod”model (*triangles*). For details, see [5]. The QMC datum (*blacked hatched circle*) for a HIGFET, $r_{\mathrm{s}} = 5$, is from [22]

affecting high densities and thick layers. A most interesting observation is that the transition to a ferromagnetic state which occurs at $r_s \simeq 26-27$ in ideal 2D, is pushed to higher r_s as the thickness w is increased.

Multi-Component 2D Layers

The two most important multi-component systems are 2D layers found in Si/SiO$_2$ interfaces, and in graphene monolayers.

Two-Valley System in Silicon-Silica Interfaces

Bulk Si has six equivalent valleys. However, when SiO$_2$ layers are grown on the (001) surface, with z the growth direction, the 2D layer occupies two degenerate layers in the (001) plane. This is a 4-component fluid, with two spin states and two valley states. Thus using an index $\nu = 1, 2, 3, 4$ where the first two are for spin, and the last two are for the valleys, there are 10 PDFS in the upper triangle of the matrix $g_{\nu\eta}(r)$. Only five of these are independent even for $\zeta \neq 0$. Four of these 10 are diagonal and contribute to both exchange and correlation, where as the off diagonals contribute *only* to the correlation energy E_c. This contrasts with the usual single valley system where there are two diagonals, and one off-diagonal PDF in the matrix of PDFs. Thus correlation effects dominate over exchange effects in multi-valley systems, and they show less tendency to spin-polarization effects which are driven by exchange effects. Thus there is no spontaneous spin-polarized state in the 2D-two valley system. However, the large increase in the correlation energy has been argued to be responsible for a rapid increase in the effective mass of electrons in Si MOSFETS as the density n decreases towards $\simeq 8 \times 10^{10}$ electrons/cm^2 [8,15]. However, this is found to be sensitive to slight changes in the CHNC parametrization.

Quantum Monte Carlo results of $E_{xc}(r_s)$ at $\zeta = 0, 1$ and CHNC results (at many values of ζ) are available, and are in excellent agreement with each other. If the total electron number density per unit area in the 2D layer is n in atomic units, then the overall $r_s = 1/\sqrt{\pi n}$, while the valley density, $n/2$ yields an $r_{sv} = \sqrt{2} r_s$. This r_{sv} and the corresponding k_{Fv} are used in constructing the non-interacting PDFs $g^0_{\nu\eta}$ from which the exchange energy can be calculate. Clearly, this is $2\epsilon_x(r_{sv})$, where ϵ_x is the usual one-valley formula for E_x/N .

The correlation energy cannot be completely mapped into results based on the 1-valley system. Thus E_c is determined via a coupling constant integration over the 10 PDFs, obtained from QMC or CHNC, using the adiabatic connection formula. Detailed results and numerical fits are available in [8].

Graphene: 2D Two-Valley System on a Honey-Comb Lattice

The electrons in the two-valley 2D system of Si/SiO$_2$ interfaces carry a valley index, but the single-particle wavefunctions are simple plane waves. In graphene, the two-component wavefunctions of massless fermions have a “chiral” character, since they carry phase angles ϕ_k, where $\vec{k}$ is the momentum vector in the 2D plane. The valence and conduction bands touch at the $\mathbf{K}, \mathbf{K}'$ points, and we need a band index $b = 1, -1$, to indicate the conduction and valence bands. Thus, we may have simultaneous electron and hole occupations, leading to an eight-component problem. Although the ϵ_x can be calculated explicitly, the calculation of the ϵ_c becomes very demanding. However, as the coupling constant g is independent of of the electron density, being $\simeq 2.7/\varepsilon$, the interactions are weak. Also, for pure electron doping or pure hole doping, the ϵ_c of the 4-component system in Si/SiO$_2$ at the same coupling strength as g (i.e, $r_s = g$) may be used as an estimate. Hence we focus on E_x and introduce a unit of energy $E_u = v_F k_c$ Here v_F is the Fermi velocity of the massless fermions, and k_c is a cutoff wavenumber chosen so that the number of states in the Brillouin zone is conserved, that is, if A_0 is the area per carbon atom, $k_c^2 = 4\pi(1/A_0)$. Then the exchange energy per Carbon atom can be written as:

$$E_x/E_u = -\frac{A_0 g^0/k_c}{(2\pi)^2}\frac{1}{4}\sum_{b_1,b_2,\sigma}\int_0^{2\pi} d\theta\, dk\, dp \times kp\frac{1+b_1 b_2\cos(\theta)}{|\mathbf{k}-\mathbf{p}|} n_{b_1,\sigma}(k) n_{b_2,\sigma}(p)\,. \quad (9)$$

In the above $g^0 = 2.7/\varepsilon$. A special feature, not found in the usual (non-chiral) 2D system is the occurrence of a term which depends on θ, the scattering angle. Because of this modulation of the interaction potential, the exchange energy in graphene is weaker than in a normal 2D layer with the same coupling strength. The above expressions for the exchange energy may be expressed [19] in terms of integrals over PDFs involving Bessel and Struve functions, or as elliptic elliptic integrals which depend on ζ and the fractional doping. Calculations of the total energy, (i.e, kinetic+E_{xc}) for graphene layers doped with electrons, holes or both, *do not* support a transition to a spin-polarized ground state. This is not surprising since the strength of the coulomb interaction is $\sim 2.7/\varepsilon$, i. e., weak, given that the spin transition in usual 2D systems occur for $r_s \sim 26$.

Exchange and Correlation in a Magnetic Field

Electrons respond to an external magnetic field by forming Landau Levels. If B is the magnetic field in the z-direction, then the energy spectrum is that of a harmonic oscillator, where the cyclotron frequency ω_c and the magnetic length ℓ are given by:

$$\omega_c = |e|B/(m_{bc}) \,, \quad \ell = \{\hbar c/(|e|B)\}^{1/2} \,. \tag{10}$$

Here m_b is the band mass and c is the velocity of light. The Harmonic oscillator energy levels with index η are:

$$\epsilon_\eta = (\eta + 1/2)\hbar\omega_c + \epsilon_{sb} \tag{11}$$

contains the lowest subband energy ϵ_{sb} of z-motion, and this may be taken as the zero of energy. The momentum states of the 2D system collapse into degenerate states, with each Landau level having $N_L = eBA/hc$ states, where A is the area of the system. If there are N electrons in A, the *filling factor* $\nu = N/N_L = n/(2\pi\ell^2)$. The integer part η of ν gives the number of *filled* Landau levels, while $\nu - \eta$ is the part of the highest landau level which is partially field. The filled Landau levels are like "inert shells" of atoms, while the partially filed part is polarizable and accounts for the more interesting aspects of exchange and correlation among electrons in magnetic fields. The total filling factor $\nu = \nu_1 + \nu_2$ where ν_i is for the ith spin species.

Since we are interested in ϵ_{xc}, we need the PDFs for use in Eq. (1). Thus it is convenient to work in the symmetric gauge where the eigenstates are functions of the radial distance r and the angle ϕ in the 2D plane, having radial and angular momentum quantum numbers n_r and m. Then, with $\hbar = 1$,

$$\epsilon_{n_r,m} = \{n_r + (|m| - m)/2 + 1/2\}\omega_c \,. \tag{12}$$

When the field is weak, there are a large number of Landau levels, and the usual 2D formulae for exchange and correlation can be used. This can be improved by using the summations over the density of states instead of integrations over k_x, k_y. A discussion may be found in, e. g., [4]. The high magnetic field limit, where only a few occupied Landau levels is of great importance. The integer and fractional QHE are found in this 'high-field' regime. If the number of full Landau levels of spin σ is ν_σ, the non-interacting PDF $g^0(r)$ is:

$$g^0(r) = 1 - \frac{e^{-x/2}}{\nu^2} \sum_\sigma \left[L^1_{\nu_\sigma - 1}(x^2/2)\right]^2 \tag{13}$$

where $x = r/\ell$, and $L_n{}^1$ is an associated Laguerre polynomial. The spin summation is unimportant in high fields, when the LLL is fully polarized by the field. ϵ_x can be evaluated from Eq. (2). The 2D Coulomb interaction may contain finite-layer thickness corrections. The effect of the magnetic field is totally contained in $e_{mag} = e^2/(\varepsilon\ell)$ and forms a natural energy unit for this system. The exchange energy ϵ_ν (in e_{mag} units), is $-\sqrt{\pi/8}$ at $\nu = 1$, and increases almost linearly to $2.35\epsilon_1$ at $\nu = 6$.

The correlation energy $\epsilon_c(r_s, \nu)$ of filled Landau levels can be evaluated using the adiabatic connection formula and the interacting PDFs. However, only an RPA evaluation is available, and $\epsilon_c(r_s, \nu)$ is very close to the zero. Hence the $B = 0$ value of ϵ_c is a good approximation.

The highest occupied Landau level is partially filled if $\nu_L = \nu - \eta$ is nonzero. The correlations in partially field states are clearly manifest in the high field limit when the lowest Landau level is fractionally occupied. Here ϵ_c has to be calculated from Quantum Monte Carlo simulations, or from methods based on the plasma analogy, via an ansatz for the many particle wavefunction. The case $\nu_L = 1/3$ and other "odd-integer" fractional-QHE states were elegantly explained by Laughlin [4], leading to a Nobel prize. This can be extended to other fractions within the plasma analogy to construct PDFs and obtain ϵ_c via the adiabatic connection formula [17]. Quantum simulations as well as composite-Fermion models have now largely superseded such methods. The literature on the quantum Hall effect [4] should be consulted for more details.

Exchange and Correlation at Finite Temperatures

The first-order (exchange) free energy F_x consists of F_x^i, where i denotes the two spin species. At $T = 0$ these reduce to the exchange energies:

$$E_i^x/n = \frac{8}{3\sqrt{\pi}} n_i^{1/2} \,. \tag{14}$$

Here $n_1 = n(1+\zeta)/2$, and $n_2 = n(1-\zeta)/2$. Then the exchange energy per particle at $T = 0$, i. e., the total E_x/n becomes

$$E_x/n = \left(E_1^x + E_2^x\right)/n = -\frac{8}{3\pi r_s}\left[x_1^{3/2} + x_2^{3/2}\right] \tag{15}$$

where x_1 and x_2 are the fractional compositions $(1 \pm \zeta)/2$ of the two spin species.

We define a reduced temperature $t = T/E_F$, $E_F = \pi n$, and the species-dependent reduced chemical potentials μ_i^0/T by η_i, reduced temperatures $t_1 = t/(1+\zeta)$ and $t_2 = t/(1-\zeta)$, based on the two Fermi energies E_{F1} and E_{F2} which are $E_F(1 \pm \zeta)$. Then we have:

$$F_i^x/E_i^x = \frac{3}{16} t_i^{3/2} \int_{-\infty}^{\eta_i} \frac{I^2_{-1/2}(u)\mathrm{d}u}{(\eta_i - u)^{1/2}} \,. \tag{16}$$

The $I_{-1/2}$ is the Fermi integral defined as usual:

$$I_\nu(z) = \int_0^\infty \frac{\mathrm{d}x x^\nu}{1+\mathrm{e}^{x-z}} \,. \tag{17}$$

The η_i are given by

$$\eta_i = \log(\mathrm{e}^{1/t_i} - 1) \,. \tag{18}$$

In the paramagnetic case Eq. (16) reduces to the result given by Isihara et al. [13]. For small values of t, the exchange energy is of the form,

$$E_\mathrm{x}(r_\mathrm{s}, t) = E_\mathrm{x}(r_\mathrm{s}, 0)[1 + (\pi^2/16)t^2 \log(t) - 0.56736t^2 + \ldots] \,. \tag{19}$$

The total exchange free energy is $F_\mathrm{x} = \Sigma F^x_i$.

A real-space formulation of $F_\mathrm{x} = F^x_1 + F^x_2$ using the zeroth-order PDFs fits naturally with the approach of our study. Thus

$$F_\mathrm{x}/n = n\int \frac{2\pi r\mathrm{d}r}{r}\sum_{i<j} h^0_{ij}(r) \,. \tag{20}$$

Here $h^0_{ij}(r) = g^0_{ij}(r) - 1$. In the non-interacting system at temperature T, the antiparallel h^0_{12}, viz., $g^0_{12}(r, T) - 1$, is zero while

$$h^0_{11}(\mathbf{r}) = -\frac{1}{n_i^2}\Sigma_{\mathbf{k}_1,\mathbf{k}_2} n(k_1)n(k_2)\mathrm{e}^{i(\mathbf{k}_1-\mathbf{k}_2)\cdot\mathbf{r}} \;=\; -[f(r)]^2 \,.$$

Here k, r are 2D vectors and $n(k)$ is the Fermi occupation number at the temperature T. At $T = 0$ $f(r) = 2J_1(k_i r)/k_r$ where $J_1(x)$ is a Bessel function.

The exchange free energy is a universal function $F_\mathrm{x}(t)/E_\mathrm{x}$, for arbitrary ζ. That is, the same function applies to any component, on using the reduced Fermi temperature of the spin species. The total F_x is the sum of both spin terms. A parametrized fit is:

$$F^x_i(t,\zeta)/E^x_i(\zeta) = \frac{1 + C_1 t_i + C_2 t_i^2}{1 + C_3 t_i + C_4 t_i^2}\tanh(1/\sqrt{t_i}) \,. \tag{21}$$

The fit coefficients C_i are 3.27603, 4.81484, 3.33100, 6.51436. The temperature t_i is $t/(1 \pm \zeta)$, appropriate to the spin polarization. The exchange effects in the 2DES decay more slowly with temperature than in the 3D case where a $\tanh(1/t)$ factor appears in Eq. 3.2 of [20]. The above form does not explicitly contain the low-temperature logarithmic term [12], but it reproduces the value of 0.99382 at $t = 0.05$, while the numerical integration gives 0.9939497. Similarly, at $t = 1$, 10 and 30 the fit (integral) returns 0.63839 (0.63839), 0.22999 (0.22990), and 0.13421 (0.13410) respectively.

If the 2D layer has a finite thickness, E_x is reduced. For numerical procedures and parametrized forms, the reader is referred to [5]. The correlation energy $\epsilon_\mathrm{c}(r_\mathrm{s}, \zeta, T)$ can be easily calculated using the PDFs from CHNC [7], but they have not been presented in a parametrized form.

Conclusion

An understanding of electron-exchange and correlation, i. e., many-body effects, is crucial to the design of quantum-well lasers, field-effect transistors, spintronics and other areas of nanostructure technology. We currently have excellent results for uniform electron systems at zero magnetic fields. The case of moderate magnetic fields is poorly understood. The very high-field case has progressed with our understanding of the fractional and integer quantum Hall effects.

Bibliography

1. Abramovitz M, Stegun IA (1965) Handbook of mathematical functions. Dover, New York
2. Ando T, Fowler B, Stern F (1982) Rev Mod Phys 54:437
3. Attaccalite C, Moroni S, Gori-Giorgi P, Bachelet GB (2002) Phys Rev Lett 88:256601; (2003) Erratum 91:109902
4. Chakraborty T (1996) The fractional quantum hall effect. Springer, New York
5. Dharma-wardana MWC (2005) Phys Rev B 72:125339
6. Dharma-wardana MWC, Perrot F (2002) Phys Rev B 66:14110
7. Dharma-wardana MWC, Perrot F (2003) Phys Rev Lett 90:136601
8. Dharma-wardana MWC, Perrot F (2004) Phys Rev B 70:035308
9. Geim AK, Novoselov KS (2007) Nat Mater 6:183–191
10. Gori-Giorgi P, Perdew JP (2001) Phys Rev B 64:155102
11. Harris J, Jones RO (1974) J Phys F 4:1170
12. Hong S, Mahan GD (1995) Phys Rev B 52:7860
13. Isihara A, Toyoda T (1980) Phys Rev B 21:p3358
14. Kim Y-H, Lee I-H, Nagaraja S, Leburton J-P, Hood RQ, Martin RM (2000) Phys Rev B 61:5202
15. Kravchenko SV, Sarachik MP (2004) Rep Prog Phys 67:1
16. Liu HC, Song CY, Wasilewski ZR, Springthorpe AJ, Cao JC, Dharma-wardana C, Aers GC, Lockwood DJ, Gupta JA (2003) Phys Rev Lett 90:077402
17. MacDonald AH, Aers GC, Dharma-wardana MWC (1985) Phys Rev B 31:5529
18. Mahan GD (1981) Many-particle physics. Plenum, New York
19. Peres NMR, Guinea F, Castro Neto AH (2007) Phys Rev 72:174406; Dharma-wardana MWC (2007) Ibid 75:075427
20. Perrot F, Dharma-wardana MWC (1984) Phys Rev A 30:2619
21. Perrot F, Dharma-wardana MWC (2001) Phys Rev Lett 87:206404
22. Tan YW, Zhu J, Stormer HL, Pfeiffer LN, Baldwin KN, West KW (2005) cond-matt/0412260, Phys Rev Lett 94:16405; De Palo S,

Botti M, Moroni S, Senatore G (2005) Phys Rev Lett 94:226405; Vakili K, Shkolnikov YP, Tutuc E, DePoortere EP, Shayegan M (2004) Phys Rev Lett 92:226401

23. Tanatar B, Ceperley DM (1989) Phys Rev B 39:5005
24. chandre.dharma-wardana@nrc-cnrc.gc.ca, http://babylon.phy.nrc.ca/ims/qp/chandre/

Spin Dynamics in Disordered Solids

Fridrikh S. Dzheparov
Institute for Theoretical and Experimental Physics,
Moscow, Russia

Article Outline

Glossary

Averaging Two kinds of averaging are necessary for any observable $\hat{Q}$: a standard quantum mechanical one $Q = \langle\hat{Q}\rangle = \mathrm{Tr}(\hat{Q}\rho)$, where ρ is the density matrix, and consequent averaging over all possible positions of particles in disordered media $\langle Q\rangle_c = \langle\langle\hat{Q}\rangle\rangle_c$.

Continuum media approximationj Positions of particles forming the disordered solids can be considered as a subset (impurity sites) of crystal lattice sites, randomly distributed on the lattice with a small probability $c \ll 1$ to find a given site occupied. Many important results can be received in continuum media approximation (CMA) when prime cell volume $\Omega_c \to 0$ together with impurity concentration $c \to 0$ at a fixed value of impurity density $n = c/\Omega_c$.

Disordered solids Statically disordered media are considered, this means that constituent particles (atoms, ions and so on) do not participate in significant translational motions during the relaxation time under discussion, and therefore their positions are fixed (frozen) in the main approximation.

High-temperature approximation High-temperature approximation (HTA) in spin dynamics consists of using the simplest density matrix $\rho = 1/\,\mathrm{Tr}\,1$, corresponding to the limit of infinite temperature T of canonical Gibbs distribution $\rho_G = \exp(-H/T)/\,\mathrm{Tr}\exp(-H/T)$; here H is the Hamiltonian. Spin dynamics remains nontrivial and rich in this limit. Sometimes HTA means application of $\rho = (1 - H/T)/\,\mathrm{Tr}\,1$.

Local field See "Secular part of dipole–dipole interactions".

Local frequency The frequency $\omega_{\mathrm{loc}} = \gamma H_{\mathrm{loc}}$ of rotation of the spin in local field H_{loc}.

Secular part of dipole–dipole interactions As a rule any spin is considered as subjected to an external static magnetic field $\mathbf{H}_0 = H_0\mathbf{n}_z$ (directed along the z-axis) and local field, produced by dipole magnetic moments of surrounding spins. The Hamiltonian of the dipole–dipole interaction of spins $\mathbf{I}_i$ and $\mathbf{I}_j$ is of the form $H_d^{(0)}(i,j) = r_{ij}^{-3}((\mathbf{m}_i\mathbf{m}_j) - 3(\mathbf{n}_{ij}\mathbf{m}_i)(3\mathbf{n}_{ij}\mathbf{m}_j))$, where $\mathbf{r}_{ij} = \mathbf{n}_{ij}r_{ij} = \mathbf{r}_i - \mathbf{r}_j$, and $\mathbf{r}_i$ is the position of spin "i" having magnetic moment $\mathbf{m}_i = \hbar\gamma_i I_i$. If $H_0 \gg H_{\mathrm{loc}}$, where H_{loc} is the mean square field produced at any spin by surrounding spins (local field), then the Hamiltonian $H_d^{(0)}(i,j)$ can be substituted by the so-called secular dipole–dipole Hamiltonian $H_d(i,j) = \frac{1}{2r_{ij}^3}(1-3(\mathbf{n}_{ij}\mathbf{n}_z)^2)(\mathbf{m}_i\mathbf{m}_j - 3(\mathbf{m}_i\mathbf{n}_z)(\mathbf{m}_j\mathbf{n}_z))$. The accuracy of the substitution is not less than $\sim H_{\mathrm{loc}}/H_0$.

Spin dynamics Spin dynamics is considered as a time evolution of correlation functions directly connected with measurable quantities of paramagnetic samples.

Units As a rule $\hbar = 1$ is supposed. Part of the equations contains $\hbar$ written explicitly.

Definition of the Subject

The main specifics of the theory of spin dynamics in disordered solids result from the fact that calculation of observable values must start from the solution of the equation of motion, and then they should be averaged over random distribution of spins in the sample. Nominally any problem in statistical physics looks analogous, but the content of existing text books contains much more simple bypass methods to achieve the results. An important bypass class, determined in the preceding century, consists of problems that are equivalent to the motion of weak (or seldom) interacting (quasi)particles in translational invariant media. The basis of this solution is formed by the Boltzmann equation (devised in the 19th century) and by methods of deriving hydrodynamic equations based on this foundation [1]. Other important advancements in physical kinetics are connected with the invention of a projection technique by Nakajima–Zwanzig for deriving various master equations and with the development of effective approximations for corresponding memory functions [1,2]. Both these bypass methods produce satisfactory predictions if the system has such small parameters as, for example, the ratio of collision duration to time between collisions, or ratio of memory time to time of variation of substantial observables (the observables whereon the system evolution is projected). Similar small parameters do not exist in disordered solids after a beginning stage, or they are totally absent. Therefore, other more refined methods are necessary to predict experimental results. They are partially presented below.

Introduction

The first correct solutions of the problems of disordered media kinetics, which are directly connected with the spin dynamics, were obtained by Forster [3] and Anderson [4].

Forster's problem (in application to spin kinetics) describes the relaxation of impurity nuclei via paramagnetic impurities in the absence of nuclear spin diffusion. The evolution of the polarization $p_j(t) = \langle I_j^z(t) \rangle$ of the jth nucleus, placed at site $\mathbf{r}_j$, and having the spin I, is described by the kinetic equation

$$\dot{p}_j = -\sum\nolimits_{a=1}^{N} \nu_{aj} p_j = -\sum\nolimits_{\mathbf{r}} n_{\mathbf{r}} \nu_{\mathbf{r}j} p_j \,, \quad p_j(t=0) = 1 \,. \tag{1}$$

Here ν_{aj} is the depolarization rate under influence of the ath paramagnetic center (acceptor), N is the total number of acceptors in the sample, $\nu_{\mathbf{r}j} = \nu_{aj}(\mathbf{r}_a = \mathbf{r})$, and $n_{\mathbf{r}}$ is the occupation number of the site $\mathbf{r}$ by an acceptor ($n_{\mathbf{r}} = 1\ (0)$ if the site $\mathbf{r}$ is (not) occupied by an acceptor). Occupations of different sites will be assumed to be independent and having no dependence on $\mathbf{r}$ (with a small probability of occupation $c \ll 1$ as a rule):

$$\begin{aligned} &\langle n_{\mathbf{r}} \rangle_c = c \,, \quad \langle n_{\mathbf{r}} n_{\mathbf{x}} \rangle_c = c\delta_{\mathbf{rx}} + c^2(1 - \delta_{\mathbf{rx}}) \,, \\ &\left\langle \prod\nolimits_{j=1}^{\prime m} n_{\mathbf{r}_j} \right\rangle_c = \prod\nolimits_{j=1}^{\prime m} \langle n_{\mathbf{r}_j} \rangle_c = c^m \,. \end{aligned} \tag{2}$$

All $\mathbf{r}_j$ are different in the last relation. Coincidence in indexes can be treated using the identity $n_{\mathbf{r}}^2 = n_{\mathbf{r}}$. The problem consists of calculation of the observable polarization, averaged over all possible positions of acceptors in the sample. Ensemble averaging can be used for macroscopic samples: $p(t) = \langle p_j(t) \rangle_c$. Occupation number representation gives the simplest and general solution of the last problem [5]. Indeed

$$\begin{aligned} p(t) &= \langle p_j(t) \rangle_c = \left\langle \exp\left(-\sum\nolimits_{\mathbf{r}} n_{\mathbf{r}} \nu_{\mathbf{r}j} t\right) \right\rangle_c \\ &= \prod\nolimits_{\mathbf{r}} \langle \exp(-n_{\mathbf{r}} \nu_{\mathbf{r}j} t) \rangle_c \\ &= \prod\nolimits_{\mathbf{r}} \langle 1 + n_{\mathbf{r}} [\exp(-\nu_{\mathbf{r}j} t) - 1] \rangle_c \\ &= \prod\nolimits_{\mathbf{r}} \langle 1 + c [\exp(-\nu_{\mathbf{r}j} t) - 1] \rangle_c \\ &= \exp\left\{ \sum\nolimits_{\mathbf{r}} \ln [1 + c(\exp(-\nu_{\mathbf{r}j} t) - 1)] \right\} \,. \end{aligned} \tag{3}$$

The identity $f(n_{\mathbf{r}}) = f(0) + n_{\mathbf{r}}(f(1) - f(0))$ is applied here. It is valid for any realistic function $f(x)$. The relation (3) is exact for any c.

The typical depolarization rate is of the form $\nu_{aj} = \nu_0 r_0^6 \chi(\mathbf{n}_{aj})/r_{aj}^6$, where ν_0 scales the transport at minimal distance r_0, $\mathbf{n}_{aj} = \mathbf{r}_{aj}/r_{aj}$, $\mathbf{r}_{aj} = \mathbf{r}_a - \mathbf{r}_j$, $\chi(\mathbf{n}) = |Y_{21}(\mathbf{n})|^2$, and $Y_{lm}(\mathbf{n})$ is the spherical harmonics. Additional simplification is possible for the continuum media approximation (CMA), when $c \to 0$, but $p(0) - p(t) \neq 0$:

$$\begin{aligned} p(t, c \to 0) &= \exp\left(n \int \mathrm{d}^d r \left(\mathrm{e}^{-\nu_{\mathbf{r}j} t} - 1 \right) \right) \\ &= \exp\left(-(\beta_F t)^{d/6} \right) \,, \quad \beta_F \propto n^{6/d} \nu_0 r_0^6 \,. \end{aligned} \tag{4}$$

Here arbitrary spatial dimension d is considered, and impurity density $n = c/\Omega_{\mathrm{c}}$ is introduced together with the prime cell volume Ω_c.

The most important properties of Forster's result (4) consist of replacement of the simple exponential kinetics by a slower law, and in proportionality of β_F to transfer rate at average distance $\bar{r} = n^{-1/d}$: $\beta_F = \mathrm{const} \cdot \nu_{\mathbf{r}j}(|\mathbf{r} - \mathbf{r}_j| = \bar{r})$, where const doesn't depend on concentration.

Anderson introduced the model to calculate the EPR line form function $g(\omega) = \int_{-\infty}^{\infty} \frac{\mathrm{d}t}{2\pi} \mathrm{e}^{-i\omega t} F(t)$ and free induction decay $F(t) = \langle\langle S_+(t) S_- \rangle\rangle_c \,/\, \langle\langle S_+ S_- \rangle\rangle_c$ for the system of equivalent spins $S_j = 1/2$, randomly distributed inside a sample and having dipole–dipole interaction. The model consists of substitution of exact secular interaction

$$\begin{aligned} H &= \frac{1}{2} \sum_{ij} b_{ij} \left(S_i^z S_j^z - \frac{1}{3} \mathbf{S}_i \mathbf{S}_j \right) \\ &= \frac{1}{2} \sum_{\mathbf{r},\mathbf{q}} n_{\mathbf{q}} n_{\mathbf{r}} b_{\mathbf{rq}} \left(S_{\mathbf{r}}^z S_{\mathbf{q}}^z - \frac{1}{3} \mathbf{S}_{\mathbf{r}} \mathbf{S}_{\mathbf{q}} \right) \end{aligned} \tag{5}$$

by a simpler one, depending on spin z-components only:

$$H_A = \frac{1}{2} \sum_{\mathbf{r},\mathbf{q}} n_{\mathbf{q}} n_{\mathbf{r}} b_{\mathbf{rq}} S_{\mathbf{r}}^z S_{\mathbf{q}}^z \,, \tag{6}$$

where $S_{\mathbf{r}}^\alpha$ is the α component of spin $\mathbf{S}_{\mathbf{r}}$ placed at site $\mathbf{r}$ and $b_{\mathbf{rq}} = b_0 r_0^3 (1 - 3\cos^2 \vartheta_{\mathbf{rq}}) / |\mathbf{r} - \mathbf{q}|^3$ is the standard dipole–dipole coefficient [2] (here and below $b_{\mathbf{rr}} = 0$). It is of great importance, that free induction signals $F_2(t) = \langle S_+(t)S_-\rangle_0 / \langle S_+S_-\rangle_0$ are the same for both Hamiltonians for two spin problem, if $S_j = 1/2$. Anderson's model has an exact solution for arbitrary concentration [6]. Indeed, the equation of motion for orthogonal spin components has the solution

$$\begin{aligned} n_{\mathbf{r}} S_{\mathbf{r}}^{+}(t) &= n_r \exp(iH_A t) S_{\mathbf{r}}^{+} \exp(-iH_A t) \\ &= n_{\mathbf{r}} S_{\mathbf{r}}^{+} \exp\left(i \sum\nolimits_{\mathbf{q}} n_{\mathbf{q}} b_{\mathbf{rq}} S_{\mathbf{q}}^{z} t \right) . \end{aligned}$$

Introducing total moment of the sample $\mathbf{S} = \sum_{\mathbf{r}} n_{\mathbf{r}} \mathbf{S}_{\mathbf{r}}$, and using standard high-temperature approximation for free induction decay (FID) we have

$$\begin{aligned} F(t) = \frac{\langle\langle S_+(t)S_-\rangle\rangle_c}{\langle\langle S_+S_-\rangle\rangle_c} &= \left\langle\left\langle \exp\left(i \sum\nolimits_{\mathbf{q}} n_{\mathbf{q}} b_{\mathbf{rq}} S_{\mathbf{q}}^{z} t \right) \right\rangle\right\rangle_c \\ &= \prod\nolimits_{\mathbf{q}} \langle \cos(n_{\mathbf{q}} b_{\mathbf{qr}} t/2) \rangle_c \\ &= \prod\nolimits_{\mathbf{q}} \left(1 + c\left(\cos\left(b_{\mathbf{qr}} t/2\right) - 1\right)\right) \\ &= \exp\left(\sum\nolimits_{\mathbf{q}} \ln\left(1 + c\left(\cos(b_{\mathbf{qr}} t/2) - 1\right)\right) \right) . \end{aligned} \tag{7}$$

Applying CMA to (7) we arrive at Anderson's result:

$$\begin{aligned} F(t, c \to 0) &= \exp\left(-n \int \mathrm{d}^d q \left(1 - \cos\left(b_{\mathbf{qr}} t/2\right)\right) \right) \\ &= \exp\left(-(D_A t)^{d/3} \right) . \end{aligned} \tag{8}$$

For three-dimensional systems the model evidently has simple Lorentz form function:

$$g(\omega) = g_A(\omega) = \frac{D_A}{\pi\left(\omega^2 + D_A^2\right)} .$$

Anderson's parameter D_A, as well as β_F, is proportional to the rate of the process taken at the average distance: $D_A \propto b_{\mathbf{qr}}(|\mathbf{q} - \mathbf{r}| = \bar{r})$. Other derivations of Eq. (8) (closer to the original Anderson's treatment fulfilled for $d = 3$) can be found in [7].

Delocalization of Nuclear Polarization in a Disordered Spin System

One of the most important generalizations of the Forster process is presented by delocalization of nuclear polarization in the system of impurity nuclei, randomly distributed in a diamagnetic matrix, when host nuclei have faster phase relaxation and flip-flop transitions then impurities. Examples of such systems: nuclei ^{6}Li in the single crystal ^{7}Li^{19}F or spins ^{107}Ag in the single crystal ^{109}Ag^{19}F. The former system is of special interest, because it is accessible for direct experimental study due to unique coincidence of g-factors of stable nuclei ^{6}Li and β-active nuclei ^{8}Li (β-nuclei) [8]. The system has simple and instructive ergodic properties as well [9]. Therefore, we will consider below a specific system (consisting of ^{6}Li nuclei in a LiF single crystal with addition of one β-nucleus ^{8}Li) where nuclear polarization transfers from initially polarized ^{8}Li nucleus to the nearest nonpolarized ^{6}Li nuclei and then migrates over other ^{6}Li nuclei and might return back to the ^{8}Li.

It was shown in [10] that modern master equation treatment leads to a description of the system by following kinetic equations, which have been formulated in [5]:

$$\frac{\partial p_{i0}}{\partial t} = -\sum\nolimits_j \left(\nu_{ji} p_{i0} - \nu_{ij} p_{j0} \right) , \quad p_{i0}(t = 0) = \delta_{i0} . \tag{9}$$

Here $p_{i0} = \langle I_i^z \rangle$ is the quantum statistical average value of the z-component (polarization) of the ith nucleus of the system ^{8}Li-^{6}Li, placed at $\mathbf{r}_i$ ($i = 0$ corresponds to ^{8}Li with a spin $I_0 = I = 2$, and $i \neq 0$ to ^{6}Li with a spin $I_i = S = 1$). The rates of polarization transfer are of the form:

$$\begin{aligned} \nu_{ji} &= \xi_j \nu_{ji}^0 r_0^6 \cdot \left(\frac{1 - 3\cos^2\theta_{ji}}{r_{ji}^3} \right)^2 , \\ \nu_{ji}^0 &= \frac{\pi}{6} S(S + 1) \left(\frac{g_i g_j \beta_n^2}{\hbar r_0^3} \right)^2 g_{ij}(\omega_{ij}) . \end{aligned} \tag{10}$$

Here $\xi_j = I_j(I_j + 1)/[S(S + 1)]$, $g_0 = g_I = 0.8267$ and $g_{i \neq 0} = g_S = 0.8220$ are g-factors of ^{8}Li and ^{6}Li correspondingly, β_n is nuclear magneton, θ_{ji} is the angle between external static field H_0 and $\mathbf{r}_{ji} = \mathbf{r}_j - \mathbf{r}_i$, $r_0 = 2.01 \cdot \sqrt{2}$ Å is the minimal distance between Li nuclei, and ω_{ij} is the difference of the Larmor frequencies. The cross-relaxation form function $g_{ij}(\omega)$ was studied in [11], and it can be taken as Gaussian $g_{ij}(\omega) \approx \exp(-\omega^2/(2M))/(2\pi M)^{1/2}$ in the main approximation. Here $M = 2M_2$, where M_2 is the second moment of the ^{8}Li NMR line. As a result ν_{ij}^0 have two values only, $\nu_{ij}^0 = \nu_0$ for transfer between ^{6}Li spins (at that $\omega_{ij} = 0$), and $\nu_{i0}^0 = \nu_1$ for transfer between ^{8}Li and ^{6}Li (with $\omega_{ij} = \Delta$).

The occupation number representation of the Eq. (9) is of the form [5]

$$\frac{\partial \tilde{P}_{\mathbf{x0}}}{\partial t} = -\sum\nolimits_{\mathbf{z}} \left(n_{\mathbf{z}} \nu_{\mathbf{zx}} \tilde{P}_{\mathbf{x0}} - n_{\mathbf{x}} \nu_{\mathbf{xz}} \tilde{P}_{\mathbf{z0}} \right) \,, \qquad \tilde{P}_{\mathbf{x0}}(t=0) = n_x \delta_{\mathbf{x0}}/c \,, \quad (11)$$

where propagator $\tilde{P}_{\mathbf{x0}}$ gives polarization of the lattice site $\mathbf{x}$ when initially the site $\mathbf{0}$ was polarized, and $\nu_{\mathbf{zx}} = \nu_{ij}(\mathbf{r}_i = \mathbf{z}, \mathbf{r}_j = \mathbf{x})$. Equation (9) is a direct consequence of Eq. (11), that is evident, if we omit all empty sites for which $n_{\mathbf{x}} = 0$ and, consequently, $\tilde{P}_{\mathbf{x0}} = 0$.

The experimentally observable value is polarization of the β-nucleus, averaged over a random distribution of ^{8}Li-^{6}Li nuclei $P_{\mathbf{00}}(t) = \langle p_{00} \rangle_c = \langle \tilde{P}_{\mathbf{00}} \rangle_c$. A calculation of such a value belongs to the problems of random walks in disordered media (RWDM), which is one of the most complex modern fields of statistical physics, which defines the actuality of this kind of study. To clarify the status of the problem we note that its solution in CMA can be connected [12] with the calculation of the path integral of the form:

$$\mathcal{P}_{\mathbf{xy}}(t) = \int_{\mathbf{q}(0)=\mathbf{x}}^{\mathbf{q}(t)=\mathbf{y}} D\mathbf{p}(\tau) D\mathbf{q}(\tau) \exp\left[i \int_{\mathbf{x}}^{\mathbf{y}} \mathbf{p} \mathrm{d}\mathbf{q} + L\left[\mathbf{q}, \mathbf{p}, t\right] \right] \quad (12)$$

$$L\left[\mathbf{q}, \mathbf{p}, t\right] = n \int \mathrm{d}^3 z \left(\mathrm{e}^{-\int_0^t \mathrm{d}\tau A^z(\mathbf{q}(\tau), \mathbf{p}(\tau))} - 1 \right) \,, \quad (13)$$

where $A^z(\mathbf{q}, \mathbf{p}) = \nu_{\mathbf{zq}} \left(1 - \mathrm{e}^{-i\mathbf{p}(\mathbf{z}-\mathbf{q})} \right)$. The representation (12) is similar to, but more complex than path integrals in the famous polaron problems (PP) [12]. The main difference consists of multi-time action L in (13) instead of two-time action in PP, in strong singularity of $\nu_{\mathbf{zq}}$ instead of the less singular kernel $1/|\mathbf{z} - \mathbf{q}|$ in PP, in additional path integral over all $\mathbf{p}(\tau)$ and in strong dependence of the infinite-fold integral (12) on the exact form of approximations by the integrals with finite multiplicity.

Field and superfield path integral representations for $P_{\mathbf{xy}}(t) = \langle \tilde{P}_{\mathbf{xy}} \rangle_c$ exist as well [13]. They demonstrate the relation of the RWDM with general problems of the modern field theory. It should be noted, that Eq. (9) and corresponding propagator $P_{\mathbf{xy}}(t)$ have wide applications in many fields of physics. For example they describe (after minimal corrections, but retaining the dipole long-range action) incoherent spatial transport of the localized electronic excitations (which is of importance in optics and biophysics) [14], and with $\ln(\nu_{ij}) \propto r_{ij}$ the same equations are used in the theory of the hopping conductivity [15], where their applications are combined with percolation theory.

The prognosis of the measurements with dipole transport has been based on the relation [10]

$$P_{\mathbf{00}}(t) = F(t) = \exp\left(-\sqrt{\beta_1 t}\right) + \xi \frac{1 - \exp\left(-\sqrt{\beta_1 t}\right)}{(\mu\beta(t+\tau))^{3/2}} \cdot \left(1 + \frac{\varphi}{\sqrt{\mu\beta(t+\tau)}} \right) \,, \quad (14)$$

where $\xi = \xi_0 = I(I+1)/[S(S+1)] = 3$, Forster parameters $\beta = \frac{256}{243} \frac{r_0^6}{\Omega_\mathrm{c}^2} \pi^3 c^2 \nu_0$ and $\beta_1 = \beta \cdot \nu_1/\nu_0$, and the limit of small c is assumed (CMA). The values $\varphi = 2.09$ and $\mu\beta\tau = 5.11$ were chosen here to construct Eq. (14) as an interpolating formula between exact results [16] of the expansion of $P_{\mathbf{00}}$ in terms of c^m (which is an expansion in powers of $(\beta t)^{m/2}$ in reality) and expansion in terms of $1/(\beta t)^{m/2}$, produced by the approximative or numerical treatment of the long time asymptotics (where dipole long ranging induces exact dependence between the first and second terms [10]). The value of μ is connected with the main values D_α of the diffusion tensor: $\mu\beta = 4\pi (c/\Omega_\mathrm{c})^{2/3} (\prod_{\alpha=1}^3 D_\alpha)^{1/3}$. It should be stressed that an analytical solution for the nature of the long time asymptotics is absent up to now, but studies of realistic models [17] and modern numerical-analytical studies [13,18,19] indicate that it is of diffusion type. The diffusion tensor is calculated now with 1% uncertainty [18,19] giving $\mu = 0.71$. As a result the relation (14) holds to within $(\beta t)^{1/2}$ at small βt, and it holds to within $(\beta t)^{-2}$ at large βt.

The applicability of the Eq. (14) (having no fitting parameters) was checked at $\beta_1 t \leq 10$ in [8] and at $\beta_1 t \leq 15$ in [20]. The last experimental results [21,22] of the ITEP group indicate that at $\beta_1 t \sim 25$ some correction is necessary. It can be introduced in simplest form as

$$P_{\mathbf{00}}(t) = F(t) G(t)$$
$$G(t) = G_{\exp} = \left(1 - \frac{\left(\frac{1}{8} + \alpha\right) \beta_1 t - u(\beta_1 t)^2}{(1 + \nu\beta t)^3} \right) \,. \quad (15)$$

Here $F(t)$ is defined in (14), $\alpha = \alpha(\Delta)$ is tabulated in [10] (at that $\alpha(\Delta \to 0) = 0.013$), and fitting parameters u and ν should be determined by experimental data. The relation (15) holds to within βt at small βt, and it holds to within $(\beta t)^{-2}$ at large βt. The fitting produces $u \approx 0.06$ and $\nu \approx 0.12$ [21]. Direct numerical simulation of the $P_{\mathbf{00}}(t)$ fulfilled in [19] for a small external magnetic field corresponding to $\beta \approx \beta_1$ justified this correction and gave more precise form and values for $F(t)$ indicating that

$0.9 \le F(t) \le 2$. New experimental results [23] have been obtained for larger magnetic fields, when $\beta - \beta_1 \propto \beta_1$. More detailed numerical simulation is expected for description of the process in this region of fields. This should take into account correlations of local fields on impurity spins [11].

The numerical simulation [13,18,19] is based on the substitution of infinite disordered media by a crystal with a large disordered elementary cell, containing $N_d \sim 1000$ impurity spins, and on checking of stability of the results with regard to variation of N_d. It starts from Eq. (9) written as

$$\frac{\partial p_{i0}}{\partial t} = -\sum_j B_{ij} p_{j0} \,, \quad B_{ij} = \delta_{ij} \sum_k \nu_{ki} - \nu_{ij} \,,$$
$$p_{i0}(t=0) = \delta_{i0} \,, \tag{16}$$

and N_d impurities randomly placed in sites of a supercell, having $N = N_d/c = N_g^3$ lattice sites, and edges $\mathbf{R}_\alpha = N_g \mathbf{b}_\alpha$, where $\mathbf{b}_\alpha$ form the basis of the matrix crystal (LiF for the system ^{8}Li-^{6}Li). Then the supercell is continued periodically to cover all space. If $N_d \to \infty$, then we go back to random media. It is very important that the initial condition remains of the correct form and that it is not continued periodically. The eigenvalue problem

$$\sum_j B_{ij}\phi_j(m) = \varepsilon_m \phi_i(m) \tag{17}$$

has Bloch's solution

$$\phi_j(m) = \exp(i\mathbf{k}\mathbf{r}_j)\chi_j(\mathbf{k},\mu) \,, \quad m = \{\mathbf{k},\mu\} \,, \tag{18}$$

where $\chi_j(\mathbf{k},\mu)$ has periodical dependence on impurity position $\mathbf{r}_j$ and $\mathbf{k}$ belongs to Brillouin zone V_B formed by all $\mathbf{k}$ satisfying the condition

$$|\mathbf{k}\mathbf{R}_\alpha| \le \pi \,.$$

Therefore, we have a finite eigenproblem for $\chi_j(\mathbf{k},\mu)$:

$$\sum_{j=0}^{N_d-1} \bar{B}_{ij}(\mathbf{k})\chi_j(\mathbf{k},\mu) = \varepsilon_{\mathbf{k},\mu}\chi_j(\mathbf{k},\mu) \,, \tag{19}$$
$$\bar{B}_{ij}(\mathbf{k}) = \sum_{n\alpha} \mathrm{e}^{-i\mathbf{k}(\mathbf{r}_i - \mathbf{r}_j - \mathbf{R}(\mathbf{n}))} B_{ij}\left(\mathbf{r}_j + \mathbf{R}(\mathbf{n})\right) \,. \tag{20}$$

Here $\mathbf{R}(\mathbf{n}) = \sum_{\alpha=1}^{3} n_\alpha \mathbf{R}_\alpha$, and $B_{ij}\left(\mathbf{r}_j + \mathbf{R}(\mathbf{n})\right)$ is B_{ij} with substitution $\mathbf{r}_j \to \mathbf{r}_j + \mathbf{R}(\mathbf{n})$. As a result

$$p_{j0}(t) = \frac{1}{V_B}\int_{V_B} \mathrm{d}^3k \exp\left(i\mathbf{k}\mathbf{r}_j\right)\left(\exp\left(-\bar{B}(\mathbf{k})t\right)\right)_{j0} \,, \tag{21}$$

$$P_{\mathbf{00}}(t) = \frac{1}{V_B}\int_{V_B} \mathrm{d}^3k \left\langle\left(\exp\left(-\bar{B}(\mathbf{k})t\right)\right)_{00}\right\rangle_c \,, \tag{22}$$

$$P_{\mathbf{x0}}(t) = \frac{c}{V_B}\int_{V_B} \mathrm{d}^3k \exp\left(i\mathbf{kx}\right)\left\langle\left(\exp\left(-\bar{B}(\mathbf{k})t\right)\right)_{10}\right\rangle_c^{(\mathbf{r}_1=\mathbf{x})} \,. \tag{23}$$

Here $\langle(\exp(-\bar{B}(\mathbf{k})t))_{10}\rangle_c^{(\mathbf{r}_1=\mathbf{x})}$ means that averaging is fulfilled under condition that site $\mathbf{x}$ is occupied by an impurity spin having number 1 (the site $\mathbf{0}$ is always occupied by spin "0").

For identical impurities

$$P_{\mathbf{00}}(t) = \frac{1}{V_B N_d}\int_{V_B} \mathrm{d}^3k \left\langle \mathrm{Tr}\left\{\exp\left(-\bar{B}(\mathbf{k})t\right)\right\}\right\rangle_c \,. \tag{24}$$

The Eqs. (22–24) were applied for numerical simulation basing on modern programs of matrix diagonalization. The results for the ^{8}Li-^{6}Li system and for electrodipole transport of localized electronic excitations among identical impurities can be found in [19].

Nuclear Relaxation via Paramagnetic Impurities

Another important generalization of Forster's process is presented by nuclear relaxation via paramagnetic impurities, which is the main relaxation channel in isolators if nuclear spin $I = 1/2$. The fundamental studies [2,24,25,26] were based on calculation of the linear (in impurity concentration c) term of the time dependence of the sample magnetization $P(t) = 1 - cQ(t)$ (for 3d-systems) with following substitution $P(t) = \exp(-cQ(t))$. Analysis of two- and one-dimensional problems was absent whereas experiments are already aimed at fractal objects [27]. Therefore, new theory was constructed [28]. It produced the relation $P(t) = \exp(-cQ(t))$ as the main approximation, and the function $Q(t)$ is calculated for arbitrary $d \le 3$.

The process used to be described by the kinetic equation

$$\frac{\partial p(\mathbf{x},t)}{\partial t} = D\Delta p(\mathbf{x},t) - \sum\nolimits_{\mathbf{z}} n_{\mathbf{z}}\nu_{\mathbf{zx}} p(\mathbf{x},t) \,,$$
$$\nu_{\mathbf{zx}} = \frac{\nu_0 r_0^6}{|\mathbf{x}-\mathbf{z}|^6} = \frac{C}{|\mathbf{x}-\mathbf{z}|^6} \,, \tag{25}$$

with the initial condition $p(\mathbf{x},t) = p_0$. Here $p(\mathbf{x},t)$ is polarization of the nucleus, placed at the crystal site $\mathbf{x}$, D is spin-diffusion coefficient, and the angular dependence of $\nu_{\mathbf{zx}}$ is neglected together with the difference in eigenvalues of the diffusion tensor. The observable nuclear polarization (normalized to $\bar{p}(t=0)=1$) is

$$\bar{p}(t)=\frac{1}{\Omega}\int \mathrm{d}^d x\, p(\mathbf{x},t)=\langle 0\,|G(t)|\,0\rangle=\langle 0\,|\langle G(t)\rangle_c|\,0\rangle\,. \tag{26}$$

Here d is the space dimensionality, Ω is the crystal volume, the symbol $|0\rangle$ presents a vector having components $\langle \mathbf{x}|0\rangle = 1/\sqrt{\Omega}$, and the propagator $G_{\mathbf{xy}}(t)=\langle \mathbf{x}|G(t)|\mathbf{y}\rangle$ obey Eq. (25), but for initial condition $G_{\mathbf{xy}}(t=0)=\delta(\mathbf{x}-\mathbf{y})$.

Expansion of the observable $\bar{p}(t)$ in concentration powers [16] gives in first terms

$$\begin{aligned}\bar{p}(t)&=\langle 0\,|G(t)|\,0\rangle=\left\langle 0\left|G^{(0)}(t)\right|0\right\rangle\\&\quad+n\int \mathrm{d}^d r\left\langle 0\left|\left[G^{(1)}(t,\mathbf{r})-G^{(0)}(t)\right]\right|0\right\rangle+O(n^2)\\&=\exp\left(-M_0(t)(1+O(n^2))\right)\,,\end{aligned} \tag{27}$$

$$M_0(t)=n\int \mathrm{d}^d r\left\langle 0\left|\left[G^{(0)}(t)-G^{(1)}(t,\mathbf{r})\right]\right|0\right\rangle\,. \tag{28}$$

The propagator $G^{(0)}(t)$ corresponds to evolution in the absence of acceptors, and $G^{(1)}(t,\mathbf{r})$ is the propagator of the system, having only one acceptor, placed at $\mathbf{r}$. Then, using operator representation of Eq. (25), translational invariance, resolvent identity and spectral expansion, we have

$$\begin{aligned}M_0(\lambda)&=\int_0^\infty \mathrm{d}t\mathrm{e}^{-\lambda t}M_0(t)\\&=\frac{n\Omega}{\lambda}\left\langle 0\left|U_0\frac{1}{\lambda+A+U_0}\right|0\right\rangle\\&=\frac{n\Omega}{\lambda}\left\langle 0\left|U_0\frac{1}{(A+U_0)(\lambda+A+U_0)}U_0\right|0\right\rangle\\&=\frac{n\Omega}{\lambda}\sum_n\frac{|\langle n\,|U_0|\,0\rangle|^2}{(\lambda+E_n)\,E_n}\,.\end{aligned} \tag{29}$$

Here $A=-D\Delta$, $\langle \mathbf{x}|U_0|\mathbf{z}\rangle=\delta_{\mathbf{xz}}\nu_{\mathbf{x0}}$, and $(A+U_0)|n\rangle=E_n|n\rangle$. This representation is too complex for direct calculations, but it is useful for different asymptotical treatments (see, for example, [8]). The most important are short time and long time studies, because a satisfactory precision can be achieved with the aid of the representation

$$M_0(t)=M_F(t)+M_1(t)\,. \tag{30}$$

Here the first term

$$M_F(t)=n\int \mathrm{d}^d x\,(1-\exp(-\nu_{\mathbf{x0}}t))=(\beta_F t)^{d/s}$$

describes the initial (Forster's) part of the relaxation, while $M_1(t)$ is the long-time asymptotic expression, calculated on the basis of the representation (30) and formally continued to arbitrary positive values of t. Here we use $\nu_{\mathbf{x0}}=\nu_0 r_0^s/x^s$ to clarify some parametrical dependencies. The results, obtained for $M_1(t)$ at arbitrary $d\le 3$, are presented in detail in Ref [28]. In particular, at long time

$$\begin{aligned}M_1(t,d=1)&=4n\sqrt{\frac{Dt}{\pi}}\,,\\M_1(t,d=2)&=\frac{4\pi Dnt}{\ln\left(Dt/b^2\right)}\,,\\M_1(t,d=3)&=4\pi Dbnt\,,\end{aligned} \tag{31}$$

where $b\propto (C/D)^{1/(s-2)}$ is a "scattering length", which incorporates all dependence on the "potential" $\nu_{\mathbf{x0}}$. It is evident, that at $d=1$ the dependence on b is absent here, and for $d=2$ it is rather weak.

In order to find an argument for regrouping (27) of the concentration expansion, we represent the propagator $G(t)$ in the form

$$\begin{aligned}G(t)&=\langle\exp\left(-(A+U)t\right)\rangle_c=\exp\left(-B(t)\right)\,,\\B(t)&=At+M(t)\,,\end{aligned} \tag{32}$$

$$U=\sum\nolimits_{\mathbf{z}} n_{\mathbf{z}}U^{\mathbf{z}}\,,\quad U^{\mathbf{z}}_{\mathbf{xq}}=\delta_{\mathbf{xq}}\nu_{\mathbf{xz}}\,. \tag{33}$$

The operator $M(t)$, which is as-yet undefined, can be written in the form $M(t)=\sum_{\mathbf{z}} M_{\mathbf{z}}(t)$ similar to the form of U in (33). It can be said that the operators $M_{\mathbf{z}}(t)$ must adequately describe the effect of acceptors in the so-called effective medium that appears upon averaging over the configurations of acceptors. It is therefore natural to assume that, on average, the propagator $G(t)$ undergoes no changes if one of the sites of the effective medium is replaced by an actual one and if the result is thereupon averaged over the distribution of acceptors; that is,

$$G(t)=\left\langle\exp\left(-At-M(t)+M^z(t)-n_z U^z t\right)\right\rangle_c\,. \tag{34}$$

Relations (32)–(34) form a closed set of nonlinear operator equations. The solution, according to [28], practically coincides with $P(t)=\exp(-M_0(t))$, if $cQ(t)=M_0(t)\sim 1$ and $c\ll 1$, and corrections are important for longer time.

It should be stressed, that (1) to clarify the influence of corrections to this solution in more detail we can calculate the next ($\propto c^2$) term of the concentration expansion, and (2) there exist physical [29] and mathematical [30,31] studies, giving the law $\log\left(1/\bar{p}(t\to\infty)\right)\propto t^{d/(d+2)}$, which is expected [32] to be valid at $\bar{p}(t)\lesssim 10^{-12}$. The asymptotics can be received by methods of field theory, but up to a preexponential multiplier only [33], therefore these methods did not produce useful results for the problem of random walks in disordered media, discussed in Sect. "Delocalization of Nuclear Polarization in a Disordered Spin System", where the exponential part of the asymptotics is absent.

Resonance Line Form Function for Magnetically Diluted Solids

The line shape and the Fourier-transform-related free induction decay (FID) belong to the most important observable values in the physics of magnetic resonance. In the study of nuclear spin systems forming a crystal lattice the first ($\propto t^2$ and $\propto t^4$) terms of the expansion of FID in powers of time carry information of high importance [2]. In the theory of the line shape of disordered (magnetically diluted) electron spin systems, the first ($\propto c$ and $\propto c^2$) terms in the expansion in powers of the concentration c of paramagnetic centers play the same role [6]. The third ($\propto c^2$) term was calculated in a recent study [34] for the first time. This consideration is particularly topical in connection with the new experiments on measuring the EPR spectra of paramagnetic impurities distributed at the solid surface [35,36]. It can be expected, that modern pulse methods will produce new possibilities for measurement of FID in such magnetically disordered nuclear systems as ^{29}Si in silicon crystals for example. Some efforts in this direction can be found in the recent article [37] and in references therein.

Let the paramagnetic centers (PCs) be randomly distributed in a d-dimensional crystal lattice with the prime-cell volume Ω_c. The free induction decay in the high-temperature approximation is given by

$$G(t) = \left\langle \langle S^+(t) S^- \rangle_0 \right\rangle_c / \left\langle \langle S^+ S^- \rangle_0 \right\rangle_c \ , \tag{35}$$

where $S^\pm = \sum_{\mathbf r} n_{\mathbf r} S^\pm_{\mathbf r}$, $S^\pm_{\mathbf r} = S^x_{\mathbf r} \pm i S^y_{\mathbf r}$, $S^+(t) = \mathrm{e}^{iHt} S^+ \mathrm{e}^{-iHt}$, $n_{\mathbf r}$ is the occupation number, $\langle\cdots\rangle_0 = \mathrm{Tr}\,(\cdots)\,/\mathrm{Tr}1$, $\langle\cdots\rangle_c$ stands for the averaging over the spatial spin distributions (over occupation numbers), and H is the secular part of dipole–dipole interactions:

$$H = \frac{3}{4}\sum_{rq} n_{\mathbf r} n_{\mathbf q} A(\mathbf r, \mathbf q)\left(S^z_{\mathbf r} S^z_{\mathbf q} - \frac{a}{3}\mathbf S_{\mathbf r} \mathbf S_{\mathbf q}\right) . \tag{36}$$

Here $A(\mathbf r, \mathbf q) = \hbar\gamma^2\left(1 - 3\cos^2\vartheta_{rq}\right) / |\mathbf r - \mathbf q|^3$, γ is the gyromagnetic ratio, and ϑ_{rq} is the angle between $\mathbf r - \mathbf q$ and external static field $\mathbf H_0$. Parameter $a = 0$ in the Anderson model and $a = 1$ for pure dipole interaction. For other a values, Hamiltonian (36) corresponds to a system with the anisotropic axisymmetric g-factor. In what follows, $S = 1/2$ for all PCs. Let us expand FID in power of $n_{\mathbf x}$ and perform configurational averaging [6]. To terms $\propto O(c^3)$ one has

$$\begin{aligned} G(t) = {} & 1 + c\sum_{\mathbf r_1}(2K_{01}(t) - 1) \\ & + \frac{c^2}{2}\sum_{\mathbf r_1 \neq \mathbf r_2}(2K_{012}(t) - 2K_{01}(t) - 2K_{02}(t) + 1) \ , \end{aligned} \tag{37}$$

and, applying CMA we have up to terms $\propto O(n^3)$

$$\begin{aligned} G(t) = {} & 1 + n\int \mathrm{d}^d r_1\,(2K_{01}(t) - 1) \\ & + \frac{n^2}{2}\int \mathrm{d}^d r_1 \mathrm{d}^d r_2\,(2K_{012}(t) - 2K_{01}(t) \\ & - 2K_{02}(t) + 1) \ , \end{aligned} \tag{38}$$

where $n = c/\Omega_c$ is the d-dimensional PC density, and

$$K_{01}(t) = \left\langle \mathrm{e}^{iH_{01}t} S_0^+ \mathrm{e}^{-iH_{01}t}(S_0^- + S_1^-)\right\rangle_0 \ ,$$

$$K_{012}(t) = \left\langle \mathrm{e}^{iH_{012}t} S_0^+ \mathrm{e}^{-iH_{012}t}(S_0^- + S_1^- + S_2^-)\right\rangle_0 \ .$$

Here, according to Eq. (36),

$$H_{ij} = \tfrac{1}{2}A_{ij}(3S_i^z S_j^z - a\mathbf S_i \mathbf S_j) , \quad H_{012} = H_{01} + H_{02} + H_{12} , \tag{39}$$

with $A_{ij} = A(\mathbf r_i, \mathbf r_j)$. The interaction $A_{ij} \propto \left|\mathbf r_i - \mathbf r_j\right|^{-3}$. Therefore, substitution of integration variables $\mathbf r_i \to t^{1/3}\mathbf r_i$ excludes time from the integrands and reveals that the mth term in (38) is $\propto (nt^{d/3})^m$. That is Eq. (38) is expanded [16] in terms of dimensionless parameter

$$(D_d t)^{d/3} = n\int \mathrm{d}^d r_1\,(1 - 2K_{01}(t)) \ , \tag{40}$$

and, hence, can be represented as

$$G(t) = 1 - (D_d t)^{\frac{d}{3}} + \frac{1}{2}\xi_d(a)(D_d t)^{\frac{2d}{3}} + O\left((D_d t)^d\right) . \tag{41}$$

The functions $\xi_d(a)$ were calculated numerically basing on the Eq. (38). It was supposed, that at $d \leq 2$ all field directions are equally probable. For this case, after averaging of (41) we have

$$\bar G(t) = 1 - (\bar D_d t)^{\frac{d}{3}} + \frac{\bar\xi_d(a)}{2}(\bar D_d t)^{\frac{2d}{3}} + O\left((\bar D_d t)^d\right) , \tag{42}$$

and

$$\begin{aligned} & \bar D_d = \beta_d n^{d/3}\gamma^2\hbar \ , \quad \beta_3 = \frac{2\pi^2}{3\sqrt 3} \ , \\ & \beta_2 = 4.647 \ , \qquad \beta_1 = 6.348 \ . \end{aligned} \tag{43}$$

It should be noticed, that $\bar{D}_3 = D_3$, and D_3 coincides with Anderson's result [4].

The results are presented in the following table:

a	0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6
ξ_3	1	1.01	1.03	1.05	1.08	1.11	1.13	1.15	1.18
$\bar{\xi}_2$	1.027	1.07	1.11	1.15	1.19	1.22	1.25	1.27	1.29
$\bar{\xi}_1$	1.062	1.16	1.20	1.25	1.28	1.32	1.35	1.37	1.40

(44)

These relations can be used directly for analysis of the wings of experimental EPR lines [34]. But treatment of the full line or FID requires regrouping of the expansion (42) in such a way to receive a physically adopted result for all t. One of the simplest approaches [38,39,40] was generalized for this aim in [34]. It introduces the most essential properties of the disordered systems into the Anderson–Weiss–Kubo (AWK) theory [41,42], which was originally developed for the description of line narrowing by motion. The result is of the form

$$\begin{aligned} G(t) &= \exp\left(-\left(2B_d^2 \int_0^t \mathrm{d}\tau (t-\tau) F(B_d\tau)\right)^{d/6}\right) , \\ F(x) &= \exp\left(-(q_d x)^{d/3}\right) . \end{aligned} \tag{45}$$

Being approximative Eq. (45) nevertheless reproduces the structure of the expansion (41), non-negativity of the resonance line shape and parameters B_d and q_d can be defined using D_d and ξ_d, that gives

$$\begin{aligned} B_3 &= D_3 , & B_2 &= \bar{D}_2 , \\ q_3 &= 3(\xi_3 - 1) , & q_2 &= \left((10/3)(\bar{\xi}_2 - 1)\right)^{3/2} . \end{aligned} \tag{46}$$

The dimensionality $d = 1$ requires more refined treatment, because system orientations near the "magic" direction $\theta = \arccos(1/\sqrt{3})$ produce a singularity $g_1(\omega \to 0) \propto \ln(1/\omega)$.

In application to pure dipole interaction with $a = 1$ the Eqs. (46) and (44) give $q_3 = 0.33 \approx 1/3$ and $q_2 = 0.63$. Fitting of the experimental data of [35] in the region, where the sample was considered as two-dimensional gives $q_2 = q_2^{\exp} \approx 0.05$ with significant distinctions from both $q_2 = 0$, and $q_2 = 0.63$. Additional studies are necessary to clarify the nature of this deviation from pure dipole evolution.

Application of the approach to magnetically diluted nuclear spin systems (like ^{29}Si in silicon) requires one important modification connected with the fact that typical nuclear concentration $c = 0.01 \div 0.1$ is not sufficiently small to replace lattice sums by integrals in Eq. (38). Therefore, FID for such concentrations can be represented by the Eq. (45) with substitution

$$(D_d t)^{d/3} \to \varkappa(t) = c \sum_{\mathbf{r}_1} (1 - 2K_{01}(t)) . \tag{47}$$

For example, if $d = 3$ and $a = 1$, then Eqs. (45), (46) produce

$$G(t) = \exp\left(-\left(6\left(D_3 t + 3\exp(-D_3 t/3) - 3\right)\right)^{1/2}\right) . \tag{48}$$

Applying the recipe (47) we receive the representation

$$G(t) = \exp\left(-\left(6\left(\varkappa(t) + 3\exp\left(-\varkappa(t)/3\right) - 3\right)\right)^{1/2}\right) , \tag{49}$$

which is exact at small $\varkappa(t)$ up to $O(\varkappa^2(t))$ and coincides with (48) for $\varkappa(t) \gtrsim 1$.

Saturation on the Wing of a Dipole-Broadened EPR Line and Cluster Expansions

A bright demonstration of peculiarities of spin dynamics in disordered media is realized in the shape of the hole burned on the wing of the EPR line. The EPR spectrum in dipole-broadened 3d solids has the Lorentz wing, but experimentation [43] revealed that the wing of the hole falls down exponentially contrary to expectations, that a dipole broadened line should be homogeneous. To solve this contradiction a new method of cluster expansions [39,40] was invented, which gives a constructive alternative to both concentration expansion, discussed in Sect. "Resonance Line Form Function for Magnetically Diluted Solids", and spin pockets conception [44], used in [43] for interpretation of the data.

If interaction inside a pair of spins exceeds interactions with any other spin, then the spectrum of the pair is similar to a discrete one, broadened by small interactions with surrounding spins. The pair forms a two-spin cluster (2-cluster). Analogically we can define 3- and other many-spin clusters. It is sufficient for many tasks to divide all spins into three groups: 2- and 3-clusters and all other "mass" spins. Numerical analysis [39] shows, that 2- and 3-clusters contain 51(1)% and 11(1)% of all spins correspondingly, but 2-clusters define all Lorentz wings of the EPR line. Mass spins (similar to nuclear spins in a crystal) have finite heat capacity, which is defined by interaction at average distance contrary to infinite heat capacity of a full dipole–dipole reservoir, calculated in CMA. Mass spins produce fluctuating fields, which gave the main broadening of the clusters spectra. Transitions between states of the cluster take place due to interactions with other clusters and with mass spins, and they are slow. As a result saturation at the EPR line wing induce transitions between states of the 2-cluster, and the line shape of the hole is defined by interaction of the cluster with mass spins. If this interaction is estimated within the AWK model, then any mass configuration produces exponential wings of the hole $g(\Delta) \propto \exp(-\Delta/\mu)$, where Δ is detuning of the saturating field from the cluster transition frequency, and μ is defined by magnitude of the fields produced by other spins on the cluster, and by the rate of their fluctuations. After configurational averaging the wings became $\langle g(\Delta)\rangle_c \propto \Delta^{-4}$, but averaged hole shape described the transitions for small magnitude ω_1 of the saturating field only. In a general case the observable area of the hole σ should be averaged. For large ω_1 we have $\sigma \propto \ln\left(t_p\omega_1^2/\mu\right)$, where t_p is the duration of the saturating pulse. Logarithm is a slow function, therefore $\langle\sigma\rangle_c \propto \ln\left(t_p\omega_1^2/\langle\mu\rangle_c\right)$, which was observed in [43] and was interpreted as the exponential wing of Portis's pockets. The described theory produces a microscopic picture of the phenomenon and indicates that Portis's pockets have limited relation to the problem, because they are homogeneous by definition, i. e. $g(\Delta) = \langle g(\Delta)\rangle_c$, while in the cluster theory this relation isn't fulfilled.

Future Developments

Methods described in Sect. "Delocalization of Nuclear Polarization in a Disordered Spin System" are expected to be useful for understanding of studies of delocalization of relaxed excitons in different problems which can be studied by optical methods (fluorescence and multi-wave mixing) first of all. We expect that results obtained in this section will also produce a reliable basis for development of general methods of statistical physics and path integration theory in order to reproduce and clarify these results and many other problems of nonlinear field theory.

The results of Sect. "Nuclear Relaxation via Paramagnetic Impurities" create a new regular method for treatment of a wide class of problems of disordered media. The method is similar to, but, nevertheless, different from standard coherent (effective) potential approximation. We expect that the same ideas applied to the second term of the concentration expansion can produce better understanding for many related problems.

The results of Sect. "Resonance Line Form Function for Magnetically Diluted Solids" require new experimental studies for all spatial dimensions. Checking on nuclear spin systems like ^{29}Si in silicon crystals is most desirable, because the details of interaction of nuclei are known much better then for paramagnetic centers.

Cluster expansion, described in Sect. "Saturation on the Wing of a Dipole-Broadened EPR Line and Cluster Expansions" should be useful for different problems of spectral transport in disordered solids.

The work was supported by the Russian Foundation for Basic Research, project #06-02-17088.

Bibliography

1. Balesku R (1975) Equlibrium and Non-Equilibrium Statistical Mechanics. Wiley, New York
2. Abragam A, Goldman M (1982) Nuclear Magnetism: Order and Disorder. Clarendon Press, Oxford
3. Forster T (1949) Z Naturforsch (A) 4:321
4. Anderson PW (1951) Phys Rev 82:342
5. Dzheparov FS, Lundin AA (1978) Sov Phys JETP 48:514
6. Dzheparov FS, Lundin AA, Khazanovich TN (1987) Sov Phys JETP 65:314
7. Feldman EB, Lacelle S (1996) J Chem Phys 104:2000
8. Abov YG, Bulgakov MI, Borovlev SP, Gul'ko AD, Garochkin VM, Dzheparov FS, Stepanov SV, Trostin SS, Shestopal VE (1991) Sov Phys JETP 72:534
9. Dzheparov FS (1999) JETP 89:753
10. Dzheparov FS (1991) Sov Phys JETP 72:546
11. Abov YG, Gulko AD, Dzheparov FS, Stepanov SV, Trostin SS (1995) Phys Part Nucl 26:692
12. Feynman RP (1972) Statistical mechanics. W.A. Benjamin, Massachusetts
13. Dzheparov FS, L'vov DV, Shestopal VE (1998) JETP 87:1179
14. Rieger PT, Palese SP, Miller RJD (1997) Chem Phys 221:85
15. Ziman JM (1979) Models of Disorder: the Theoretical Physics of Homogeneously Disordered Systems. Cambridge University Press, Cambridge
16. Dzheparov FS, Smelov VS, Shestopal VE (1980) JETP Lett 32:47
17. Dzheparov FS, Shestopal VE (1993) Theor Mat Phys 94:496
18. Dzheparov FS, L'vov DV, Shestopal VE (2007) J Supercond Nov Magn 20(2):175
19. Dzheparov FS (2005) JETP Lett 82(8):521
20. Dzheparov FS, Gul'ko A, Heitjans P, L'vov D, Schirmer A, Shestopal V, Stepanov S, Trostin S (2001) Physica B297:288
21. Abov YG, Gul'ko AD, Dzheparov FS (2006) Phys At Nuclei 69:1701
22. Gul'ko AD, Ermakov ON, Stepanov SV, Trostin SS (2007) J Supercond Nov Magn 20(2):169
23. Dzheparov FS, Gul'ko AD, Ermakov ON et al (2008) In: Salikhov KM (ed) Modern development of magnetic resonance, Abstracts of the International conference Zavoisky-100. Kazan, pp 47–48, Appl Magn Reson 35(3)
24. Khutsishvili G (1965) Sov Phys Usp 8:743
25. Alexandrov IV (1975) Theory of Magnetic Relaxation. Nauka, Moscow (in Russian)
26. Atsarkin VA (1980) Dynamical Nuclear Polarization. Nauka, Moscow (in Russian)
27. Tabti T, Goldman M, Jacquinot JF, Fermon C, LeGoff G (1997) J Chem Phys 107:9239
28. Dzheparov FS, Jacquinot JF, Stepanov SV (2002) Phys At Nuclei 65:2052
29. Balagurov VY, Vaks VT (1973) Sov Phys JETP 38:968
30. Pastur LA (1977) Teor Mat Fiz 32:88

31. Donsker M, Varadhan S (1975) Commun Pure Appl Math 28:525
32. Havlin S, Dishon M, Kiefer JE, Weiss GH (1984) Phys Rev Lett 53:407
33. Renn SR (1986) Nucl Phys B275(FS17):273
34. Dzheparov FS, Kaganov IV (2002) JETP Lett 75:259
35. Atsarkin VA, Vasneva GA, Demidov VV, Dzheparov FS, Odintsov BM, Clarkson RB (2000) JETP Lett 72:369
36. Atsarkin VA, Demidov VV, Vasneva GA, Dzheparov FS, Ceroke PJ, Odintsov BM, Clarkson RB (2001) J Magn Res 149:85
37. Li D, Dementyev AE, Dong Y, Ramos R, Barrett SE (2007) Phys Rev Lett 98:190401
38. Grinberg ES, Kochelaev BI, Khaliullin GG (1981) Sov Phys Solid State 23:224
39. Dzheparov FS, Henner EK (1993) JETP 77:753
40. Dzheparov FS, Kaganov IV, Khenner EK (1997) JETP 85:325
41. Anderson PW, Weiss PR (1953) Rev Mod Phys 25:269
42. Kubo R (1962) J Phys Soc Jpn 17:1100
43. Atsarkin VA, Vasneva GA, Demidov VV (1986) Sov Phys JETP 64:898
44. Portis AM (1953) Phys Rev 91:1071

Spin-Polarized Quantum Transport in Mesoscopic Conductors: Computational Concepts and Physical Phenomena

Michael Wimmer, Matthias Scheid,
Klaus Richter
Institut für Theoretische Physik, Universität Regensburg,
Regensburg, Germany

Article Outline

Glossary

Aharonov Bohm effect The magnetic flux enclosed in between propagating quantum mechanical waves shifts their relative phases as a result of the underlying electromagnetic vector potential. This gives rise to distinct oscillations in the magnetoconductance of a ring conductor.

Landauer–Büttiker formalism For phase-coherent quantum transport, the Landauer–Büttiker formalism relates the conductance of a device to the transmission probability of charge carriers.

Rashba- and Dresselhaus spin-orbit coupling Coupling of the spin degree of freedom to the orbital motion of charge carriers due to structural or bulk inversion asymmetry in semiconductors.

Ratchets Devices that convert unbiased fluctuations or perturbations into directed motion.

Spintronics Extension of charge-based electronics in metals or semiconductors by utilizing the spin degree of freedom of the charge carriers.

Definition of the Subject

Mesoscopic conductors are electronic systems of sizes in between nano- and micrometers, and often of reduced dimensionality. In the phase-coherent regime at low temperatures, the conductance of these devices is governed by quantum interference effects, such as the Aharonov–Bohm effect and conductance fluctuations as prominent examples. While first measurements of quantum charge transport date back to the 1980s, spin phenomena in mesoscopic transport have moved only recently into the focus of attention, as one branch of the field of spintronics. The interplay between quantum coherence with confinement-, disorder- or interaction-effects gives rise to a variety of unexpected spin phenomena in mesoscopic conductors and allows moreover to control and engineer the spin of the charge carriers: spin interference is often the basis for spin-valves, -filters, -switches or -pumps. Their underlying mechanisms may gain relevance on the way to possible future semiconductor-based spin devices.

A quantitative theoretical understanding of spin-dependent mesoscopic transport calls for developing efficient and flexible numerical algorithms, including matrix-reordering techniques within Green function approaches, which we will explain, review and employ.

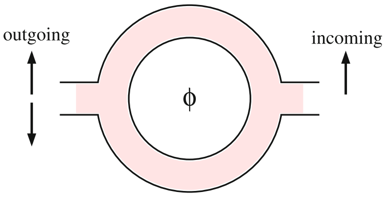
Introduction

Charge and spin transport through phase-coherent conductors of mesoscopic scales carry imprints of wave interference as predominant and characteristic features: in the simplest case of a point contact, the conductance increases stepwise with Fermi energy, reflecting the discrete number of quantized open transverse channels contributing to transport; for more complex mesoscopic systems, such as ballistic quantum dots or diffusive conductors, the conductance typically wildly fluctuates, upon varying the Fermi energy or other parameters, around its classical mean value. Among the different effects on charge transport, the Aharonov–Bohm (AB) effect represents possibly the most genuine interference phenomenon at the heart of mesoscopic physics: The magnetoconductance of a ring conductor coupled to two leads exhibits distinct sinusoidal oscillations when monitored as a function of a perpendicular magnetic field threading the ring, with a period given by the magnetic flux quantum. As the AB signal stems from interfering waves traveling through the two different arms of the ring, it requires phase coherent wave functions extending over the ring typically on micron scales [1]. Hence the AB effect is frequently being used as a tool to investigate phase coherence and dephasing mechanisms of the orbital part of the wave functions, while the spin degree of freedom was usually neglected.

With rising interest in spin-dependent transport, the interplay of the electron spin and charge degree of freedom has been exploited in a variety of spin interference devices, to be discussed below. Different types of couplings to the electron spin have been considered for spin engineering in non-magnetic conductors. On the one hand this is possible through the Zeeman coupling to an externally applied magnetic field. Non-uniform *B*-fields with spatially varying direction are being employed to achieve a tailored spin dynamics, including the possibility for guided spin evolution or triggering spin flips. On the other hand, intrinsic spin-orbit (SO) interaction proves to be relevant in spin-dependent transport. It exists in systems with bulk inversion asymmetry and/or structure inversion asymmetry, e. g. due to the vertical confinement in semiconductor heterostructures (Rashba SO coupling [2,3]).

Among the mesoscopic spin interference systems considered in the literature, ring geometries have again played an important role, opening up the field of spin-based AB physics (see [4,5] for recent accounts including overviews over the literature). This includes topics such as Berry phase-signatures in transport, both theoretically [6,7,8,9,10,11,12] and experimentally [13,14,15,16], spin-related conductance modulation [17,18,19], persistent currents [6,20], spin Hall effect [21], spin filters [22,23] and detectors [24], and spin switching mechanisms [4,10].

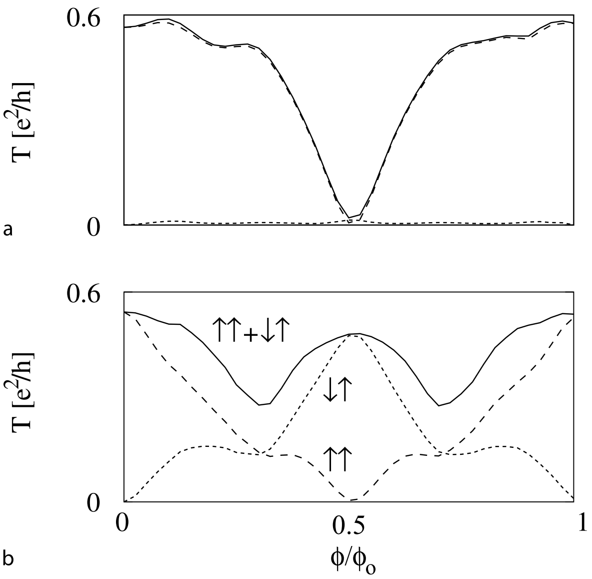
To illustrate how the spin polarization can be tuned by exploiting (orbital and spin) interference in such an AB setup, we consider as an introductory example spin switching in a two-dimensional (2D) ballistic phase-coherent ring symmetrically coupled to two single-channel leads (Fig. 1, [17]). We assume Rashba SO coupling which is relevant in conductors laterally defined on GaAs- or InAs-based two-dimensional electron gases (2DEGs). Rashba SO coupling will be defined and discussed in Sect. "Spin Filtering in Nanostructures". It can be viewed as the coupling of the spin to a fictitious in-plane magnetic field directed perpendicular to the electron momentum. Hence in a ring it points mainly in radial direction. The strength of the SO field can be tuned by an external gate voltage [25] allowing to control experimentally the spin evolution.



Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 1

Aharonov–Bohm physics with spin: Two-dimensional Aharonov–Bohm ring of mean radius r_0 used for numerical calculations of the conductance presented in Fig. 2. An additional perpendicular magnetic field B generates a flux $\phi = \pi r_0^2 B$. The *gray zone* corresponds to the region subject to a finite Rashba coupling switched adiabatically on and off in the leads (from [17])

We assume *spin-polarized* spin-up electrons entering the ring from the right (see Fig. 1). Figure 2 displays numerically computed (see Sect. "Numerical Quantum Transport") conductance traces as a function of the external magnetic flux ϕ for weak and moderate Rashba strengths. The overall conductance is presented as a solid line, and its spin-resolved components, $G^{\uparrow\uparrow}$ and $G^{\downarrow\uparrow}$, corresponding to outgoing spin-up and -down channels, are shown as dashed and dotted lines, respectively. In the weak SO coupling limit, Fig. 2a, the overall conductance (solid line) shows the usual AB oscillations of period ϕ_0 and is dominated by $G^{\uparrow\uparrow}$ (dashed line). As expected for weak spin-coupling, the spin polarization is almost conserved during transport. Interesting features appear for the case of moderate SO coupling depicted in panel (b). There, both components, $G^{\uparrow\uparrow}$ (dashed line) and $G^{\downarrow\uparrow}$ (dotted line), contribute similarly to the overall conductance (solid line). However, the spin polarization of the transmitted electrons varies as a function of the magnetic flux ϕ: $G^{\downarrow\uparrow} = 0$ at $\phi = 0$, while $G^{\uparrow\uparrow} = 0$ at $\phi = \phi_0/2$. Hence, for zero flux all transmitted carriers conserve their original (incoming) spin-orientation, while for $\phi = \phi_0/2$ the transmitted particles reverse their spin polarization. In other words: by tuning the magnetic flux from 0 to $\phi_0/2$ we can reverse the spin-polarization of transmitted particles in a controlled way. Hence the AB ring with SO coupling acts as a tunable spin-switch. This switching can be traced back to constructive or destructive AB interference [4].



Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 2
Mesoscopic Aharonov–Bohm spin interference: The conductance of spin-up polarized carriers entering a single-channel two-dimensional Aharonov–Bohm ring (see Fig. 1) is shown as a function of flux $\phi = \pi r_0^2 B$ (in units of the flux quantum $\phi_0 = hc/e$) through the ring in the presence of Rashba spin-orbit coupling. For panel a and b the scaled spin-orbit coupling strength takes the values 0.2 and 1.0, expressed as the product $\omega_R T_0/2\pi$ of the precession frequency ω_R of the spin around the effective spin-orbit magnetic field and the time T_0 for traveling of the electrons around the ring. The overall conductance (*solid line*) is split into its components $G^{\uparrow\uparrow}$ (*dashed line*) and $G^{\downarrow\uparrow}$ (*dotted line*). Note in panel b the continuous change of the spin polarization, related to $G^{\uparrow\uparrow} - G^{\downarrow\uparrow}$, with ϕ and the spin switching at $\phi = \phi_0/2$ (adapted from [17])

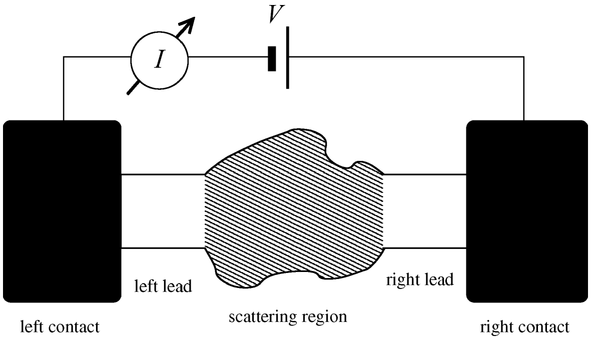
Switching a given spin polarization requires the generation of spin-polarized particles in non-magnetic mesoscopic conductors in the first place. Since spin injection from ferromagnets into a semiconductor remains problematic [26], alternative proposals have been made to achieve spin-polarized currents or spin accumulation without magnets, which we will briefly review in Sect. “Spin Filtering in Nanostructures”. Among those are the spin Hall effect [27] and, in the context of coherent mesoscopic transport, concepts for Zeeman- and SO-mediated adiabatic spin pumping and spin ratchets.

For a recent account on spin phenomena in systems of reduced dimensions see [28]; for a review on the related field of magnetization dynamics and pumping in layered magnetic heterostructures see [29].

A complete and quantitative understanding of spin phenomena in the mesoscopic realm requires computational approaches to quantum transport which also serve as reference calculations for analytical predictions usually based on model assumptions. However, also numerical approaches cannot cope with the full many-body transport problem without relying on approximations. Here we focus on mesoscopic conductors, i. e. systems with a considerable number of electrons, with strong coupling to external leads. Then, a mean-field treatment is usually justified which allows one to reduce the Hamiltonian to a single-particle problem with an effective confinement potential resulting from a combination of external and mean-field potentials.

We further consider coherent transport close to equilibrium at relatively low bias, excluding inelastic effects, such that the Landauer approach to transport is justified. However, even in this case brute-force computational approaches quickly reach their limits: Conductors at mesoscopic scales are typically characterized by extensions which are (at least in one direction) much larger than the Fermi wavelength of the charge carriers, the shortest quantum scale involved. This implies rapidly oscillating, complex and often irregular spinor wave functions extending throughout the systems which require for the quantum mechanical numerical solution either huge sets of basis functions or, in tight-binding approaches, the use of rather fine, preferentially adapted grids. The strength of the widespread tight-binding approaches for transport discussed and reviewed here lies in their flexibility and general applicability. Moreover, tight-binding transport codes can be combined with density-functional (DFT) calculations for structure and electronic properties of the nano-conductors by using the parameters computed within DFT as input. This approach is frequently applied to transport in nano- or molecular electronics.

The paper is organized as follows: In the methodological Sect. “Numerical Quantum Transport” we will first briefly summarize and provide the key relations for spin quantum transport within the Landauer framework. In Sect. “Matrix Reordering Strategies for Quantum Transport” we focus on and explain in some detail advanced computational concepts, making use of graph-theory, to implement powerful and flexible algorithms for tight-binding transport codes. The numerical strength of the codes is demonstrated for ring-type geometries showing that by efficient tight-binding implementation one can gain orders of magnitude in performance. In Sect. “Spin Filtering in Nanostructures” and “Pure Spin Current Generation” we employ these numerical schemes to address two important aspects of spin-dependent transport, namely spin filtering and generating pure spin currents. To this end we focus on laterally-confined 2D ballistic nanostructures with Rashba SO interaction. We conclude



Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 3
Schematic view of a two-point measurement setup

with an outline of future research directions in mesoscopic spin transport which we consider important.

Numerical Quantum Transport

Landauer–Büttiker Transport Theory

If the dimensions of a device get smaller than the phase coherence length l_ϕ of charge carriers, classical transport theories are not valid any more. Instead, carrier dynamics is now governed by quantum mechanics and the wave-like nature of particles becomes important. In general, the conductance/resistance of such a device does not follow Ohm's law.

Consider a two-point measurement setup as shown in Fig. 3: A scattering region is connected to large (phase-breaking) reservoirs by leads. The leads are assumed to be perfect and infinitely long to define asymptotic eigenstates $\phi_{n,\sigma}(y)\mathrm{e}^{\pm ikx}$ at energy E, where n is the quantum number of transverse confinement —also called the *channel* number —and σ is the spin index. The total scattering eigenstate $\psi_{n,\sigma}$ originating from channel n with spin σ in the left lead is, within the lead region, given by

$$\psi_{n,\sigma}(\mathbf{x}) = \begin{cases} \phi_{n,\sigma}(y)\mathrm{e}^{ikx} + \sum_m r_{m,\sigma';n,\sigma} \ \phi_{m,\sigma'}(y)\mathrm{e}^{-ikx} \\ \qquad\qquad \text{for } \mathbf{x} \text{ in left lead} \\ \sum_m t_{m,\sigma';n,\sigma} \ \phi_{m,\sigma'}(y)\mathrm{e}^{ikx} \\ \qquad\qquad \text{for } \mathbf{x} \text{ in right lead} \end{cases}, \tag{1}$$

and obeys the stationary Schrödinger equation $H\psi_{n,\sigma} = E\psi_{n,\sigma}$. The conductance G in linear response can then be calculated within the *Landauer–Büttiker* formalism [30,31,32] (for tutorials see [1,33]):

$$G = \frac{\mathrm{e}^2}{h}\sum_{n,m}\sum_{\sigma,\sigma'} T_{m,\sigma';n,\sigma} = \frac{\mathrm{e}^2}{h}\,T_{\mathrm{C}}\,, \tag{2}$$

where $T_{m,\sigma';n,\sigma} = |t_{m,\sigma';n,\sigma}|^2$ is given by the squared transmission amplitudes of the scattering states $\psi_{n,\sigma}$. The fraction e^2/h is called the conductance quantum. The *scattering matrix* $S_{m,\sigma';n,\sigma}$ is a useful definition that combines reflection and transmission amplitudes for both leads into a single matrix. In this notation the index n,σ then also contains information about the respective lead.

The problem of calculating the conductance G is thus reduced to calculating the scattering eigenstates $\psi_{n,\sigma}$. Alternatively, the scattering amplitudes $r_{m,\sigma';n,\sigma}$ and $t_{m,\sigma';n,\sigma}$ can also be derived from the retarded Greens function $G^{\mathrm{R}}(\mathbf{x},\mathbf{x}')$ of the system. The retarded Greens function for a given energy E obeys the equation

$$(E - H + \mathrm{i}\eta)\, G^{\mathrm{R}}(\mathbf{x},\mathbf{x}') = \delta(\mathbf{x}-\mathbf{x}')\,, \tag{3}$$

where H is the Hamiltonian of the system and η an infinitesimally small number. Formally this equation can be solved as

$$G^{\mathrm{R}} = (E - H + \mathrm{i}\eta)^{-1}\,. \tag{4}$$

The transmission and reflection amplitudes are then given by the *Fisher–Lee* relation [34]:

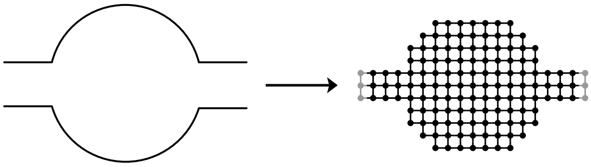
$$\begin{aligned} t_{m,\sigma';n,\sigma} &= -\mathrm{i}\hbar\sqrt{v_m v_n}\int_{C_{\mathrm{R}}}\mathrm{d}y \\ &\qquad \times \int_{C_{\mathrm{L}}}\mathrm{d}y'\phi_{m,\sigma'}(y)\, G^{\mathrm{R}}(\mathbf{x},\mathbf{x}')\,\phi_{n,\sigma}(y')\,, \end{aligned} \tag{5}$$

$$\begin{aligned} r_{m,\sigma';n,\sigma} &= \delta_{mn}\delta_{\sigma'\sigma} - \mathrm{i}\hbar\sqrt{v_m v_n}\int_{C_{\mathrm{L}}}\mathrm{d}y \\ &\qquad \times \int_{C_{\mathrm{L}}}\mathrm{d}y'\phi_{m,\sigma'}(y)\, G^{\mathrm{R}}(\mathbf{x},\mathbf{x}')\,\phi_{n,\sigma}(y')\,, \end{aligned} \tag{6}$$

where v_n denotes the velocity of channel n and the integration runs over the cross-section $C_{\mathrm{L}}(C_{\mathrm{R}})$ of the left (right) lead. Equations (5) and (6) are valid only for leads without magnetic fields and no spin-orbit interaction. Baranger and Stone [35] have extended the formalism to also account for arbitrary magnetic fields in the leads, and their description can also be applied to finite spin-orbit interaction.

Tight-binding Representation of the Hamiltonian

Except for very simple geometries, the scattering problem cannot be solved analytically. Therefore, the use of computers for a numerical solution of the scattering problem is very often the method of choice. However, the related stationary Schrödinger equation $H\psi = E\psi$ is a differential equation with continuous degrees of freedom that are difficult to treat on a computer. In general, a numerical solution is thus only attempted within a *discrete* basis set which converts the differential equation into a matrix equation.



Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 4
Discretizing a continuous region on a square grid

The method of finite differences is a very simple and yet powerful way to introduce such a discrete basis set and has been applied to the Schrödinger equation already as early as 1934 [36,37] (for an introduction see [1,33]). Here we illustrate, as an example, the application of the method for a simple one-dimensional effective mass Hamiltonian including a potential

$$H = -\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}x^2} + V(x)\,. \tag{7}$$

In the method of finite differences, space is approximated by a grid of discrete lattice points spaced equidistantly with lattice constant a. For the 2D case, usually a square grid is used as depicted in Fig. 4. Using the Taylor expansion of the wave function ψ we can write

$$\begin{aligned}\psi(x+a) &= \psi(x) + \psi'(x)a + \frac{1}{2}\psi''(x)a^2 \\ &\quad + \frac{1}{6}\psi^{(3)}(x)a^3 + \frac{1}{24}\psi^{(4)}(x)a^4 + \dots\,,\end{aligned} \tag{8}$$

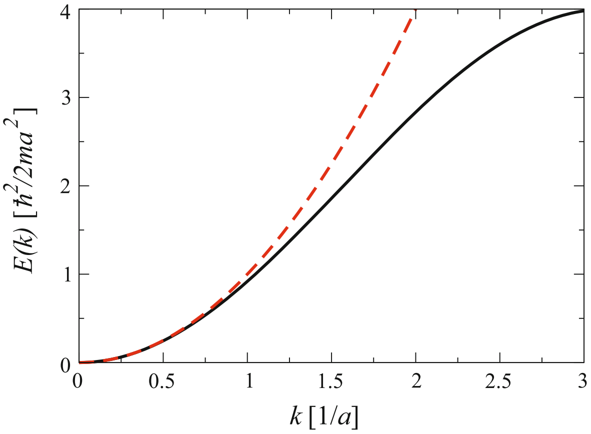
$$\begin{aligned}\psi(x-a) &= \psi(x) - \psi'(x)a + \frac{1}{2}\psi''(x)a^2 \\ &\quad - \frac{1}{6}\psi^{(3)}(x)a^3 + \frac{1}{24}\psi^{(4)}(x)a^4 + \dots\,.\end{aligned} \tag{9}$$

Adding Eq. (8) and (9), we arrive at an expression for the second derivative of the wave function in terms of values of the wave function on the grid,

$$\begin{aligned}&\frac{\mathrm{d}^2}{\mathrm{d}x^2}\psi(x) \\ &\quad = \frac{1}{a^2}\left(\psi(x+a) + \psi(x-a) - 2\psi(x)\right) + \mathcal{O}(a^2)\,,\end{aligned} \tag{10}$$

valid up to second order in the lattice spacing a. The differential equation

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\psi(x) + V(x)\psi(x) = E\psi(x) \tag{11}$$



Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 5
Energy spectrum *E*(*k*) for the continuous Schrödinger equation (*red dashed line*) and the tight-binding approximation (*black solid line*)

is thus replaced by a set of difference equations

$$\begin{aligned}-\frac{\hbar^2}{2ma^2}\left(\psi(x_{i+1}) + \psi(x_{i-1}) - 2\psi(x_i)\right) + V(x)\psi(x_i) \\ = E\psi(x_i)\,,\end{aligned} \tag{12}$$

where $x_{i\pm1} = x_i \pm a$, yielding the tight-binding representation of the Hamiltonian:

$$\begin{aligned}H = \sum_{x_i} -\frac{\hbar^2}{2ma^2}\left(|x_i\rangle\langle x_{i+1}| + \text{h.c.}\right) \\ + \left(2\frac{\hbar^2}{2ma^2} + V(x_i)\right)|x_i\rangle\langle x_i|\,.\end{aligned} \tag{13}$$

Here, $|x_i\rangle$ denotes a state localized at grid point x_i.

In principle, the quality of the finite differences approximation can be improved up to a desired precision by reducing the lattice spacing a. However, since this leads to a larger problem size, the minimum lattice spacing achievable is set by the available computing time and memory. Thus, one must keep an eye on the validity of the finite differences approximation. In Fig. 5 we show the energy spectrum $E(k)$ for the continuous one-dimensional Schrödinger equation and the tight-binding (finite differences) approximation. The tight-binding approximation only holds for $ka \ll 1$ and $E \ll \hbar^2/(2ma^2)$, and it does not make sense to consider the whole energy spectrum given by the tight-binding band width. The method of finite differences presented here can straight-forwardly be applied to more complex Hamiltonians, including for example spin-orbit interactions [17], and will be later used to calculate transport phenomena including spin in Sect. "Spin Filtering in Nanostructures" and "Pure Spin Current Generation".

A tight-binding representation of the Hamiltonian can also be obtained by applying the finite element method [38]. Furthermore tight-binding Hamiltonians are also used in treatments beyond the effective mass approximation, such as from atomic orbitals in empirical tight-binding models [39,40,41], or from orbitals of the Kohn–Sham equations within DFT [42,43,44].

Numerical Algorithms

Within the tight-binding approximation the Hamiltonian H can be represented by a matrix, even though this matrix is still infinite as the leads are infinitely long. However, the infinite matrix problem can be reduced to a finite problem by partitioning the system into three isolated parts: left lead, scattering region, and right lead. The Hamiltonian then reads

$$H = \begin{pmatrix} H_\mathrm{L} & V_\mathrm{LS} & 0 \\ V_\mathrm{SL} & H_\mathrm{S} & V_\mathrm{SR} \\ 0 & V_\mathrm{RS} & H_\mathrm{R} \end{pmatrix}\,, \tag{14}$$

where $H_{\mathrm{L(R)}}$ is the (infinite) Hamiltonian of the left (right) lead, H_S is the Hamiltonian of the scattering region and of finite size. $V_\mathrm{SL} = V_\mathrm{LS}^{+}$ and $V_\mathrm{SR} = V_\mathrm{RS}^{+}$ represent the coupling of the left and right lead to the scattering region. Since the leads are always chosen such that asymptotic eigenstates can be defined, the Hamiltonian of the isolated leads must contain some periodicity that facilitates calculating their Greens functions $g^\mathrm{R}_\mathrm{L,R}$. This can be done analytically for simple systems [1,33], for more complex situations the Greens function can be calculated numerically either by iteration [45,46] or semianalytical formulas [40,44,47]. Introducing the retarded self-energy $\Sigma^\mathrm{R} = \sum_{i=L,R} \; V_{\mathrm{S}i}\, g^\mathrm{R}_i\, V_{i\mathrm{S}}$, [1,33] the Greens function G_S of the scattering region is given by

$$G^\mathrm{R}_\mathrm{S} = \left(E - H - \Sigma^\mathrm{R}\right)^{-1}\,, \tag{15}$$

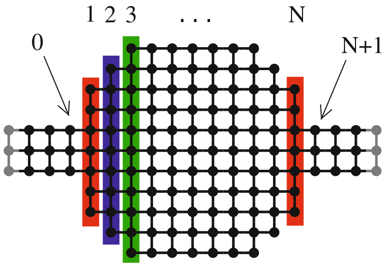
reminiscent of Eq. (4) but with an effective Hamiltonian $H + \Sigma^\mathrm{R}$.

The original infinite-dimensional problem has thus been reduced to a finite size matrix problem that can, in principle, be solved straight-forwardly on a computer. However, for any but rather small problems, the computational task of the direct inversion in Eq. (15) is prohibitive. Therefore, many algorithms make use of the *sparsity* of the Hamiltonian matrix in tight-binding representation – in particular they employ the property that this matrix can be written in block-tridiagonal form

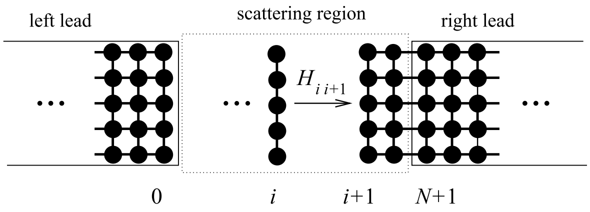
$$H = \begin{pmatrix}
\ddots & & & & & & & & \\
H_\mathrm{L} & V_\mathrm{L} & & & & & & & \\
V_\mathrm{L}^\dagger & H_\mathrm{L} & H_{01} & & \ddots & & & & \\
 & H_{10} & H_{11} & H_{12} & & 0 & & & \\
 & & H_{21} & H_{22} & H_{23} & & \ddots & & \\
 & & & H_{32} & \ddots & & & & \\
 & \ddots & & & \ddots & H_{N-1N} & & & \\
 & & 0 & & H_{NN-1} & H_{NN} & H_{NN+1} & & \\
 & & & \ddots & & H_{N+1N} & H_\mathrm{R} & V_\mathrm{R} & \\
 & & & & & & V_\mathrm{R}^\dagger & H_\mathrm{R} & \\
 & & & & & & & & \ddots
\end{pmatrix}. \tag{16}$$

Here the index L(R) denotes the blocks in the left (right) lead, $1 \ldots N$ the blocks within the scattering region, and 0, $(N+1)$ the first block in the left (right) lead. Such a form arises naturally in the method of finite differences, when grid points are grouped into vertical slices according to their x-coordinates, as shown in Fig. 6, but also applies to any other sparse tight-binding Hamiltonians.

The block-tridiagonal form of the Hamiltonian is the foundation of several quantum transport algorithms, together with the fact that, according to Eq. (5) and (6), only the blocks $G^\mathrm{R}_{N+1\,0}$ and G^R_{00} of the Greens function G^R are needed for the calculation of transmission and reflection probabilities. The transfer matrix approach applies naturally to block-tridiagonal Hamiltonians, but becomes unstable for larger systems. However, a stabilized version has been developed by Usuki et al. [48,49]. In the dec-



Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 6
Block-tridiagonal matrix form arising in the method of finite differences. Grid points with the same *x*-coordinate are placed into the same block



Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 7
Schematic view of the recursive Greens function algorithm: the system is built up adding block by block

imation technique [50,51], the Hamiltonian of the scattering region is replaced by an effective Hamiltonian between the two leads by eliminating internal degrees of freedom. The contact block reduction method [52] calculates the full Greens function of the system using a limited set of eigenstates. The recursive Greens function (RGF) technique [53,54,55] uses Dyson's equation to build up the system's Greens function block by block. It has also been adapted to Hall geometries with four terminals [56] and to calculate non-equilibrium densities [57,58]. Furthermore, the RGF algorithm has been formulated to be suitable for parallel computing [59], and the modular recursive Greens function (MRGF) method is an extension to take advantage of special geometries, such as circles or rectangles [60,61].

Here we give a simple form of the RGF algorithm as described in [33,55]. The RGF method makes use of Dyson's equation

$$G = G_0 + G_0 V G \tag{17}$$

to add successively blocks to the system, as depicted in Fig. 7. Let $G^{\mathrm{R},(i)}$ denote the Greens function for the system containing all blocks $\geq i$. Then, at energy E, the Greens function $G^{\mathrm{R},(i)}$ is related to $G^{\mathrm{R},(i+1)}$ by

$$G^{\mathrm{R},(i)}_{i\,i} = \left(E - H_{i\,i} - H_{i\,i+1}\, G^{\mathrm{R},(i+1)}_{i+1\,i+1}\, H_{i+1\,i}\right)^{-1} \tag{18}$$

and

$$G^{\mathrm{R},(i)}_{N+1\,i} = G^{\mathrm{R},(i+1)}_{N+1\,i+1}\, H_{i+1\,i}\, G^{\mathrm{R},(i)}_{i\,i}\ . \tag{19}$$

Starting from $G^{\mathrm{R},(N+1)}_{N+1\,N+1} = g^{\mathrm{R}}_{\mathrm{R}}$, the surface Greens function of the right lead, N slices are added recursively, until $G^{\mathrm{R},(1)}$ has been calculated. The blocks of the Greens function of the full system necessary for transport are then given by

$$G^{\mathrm{R}}_{00} = \left(\left(g^{\mathrm{R}}_{\mathrm{L}}\right)^{-1} - H_{01}\, G^{\mathrm{R},(1)}_{11}\, H_{10}\right)^{-1} \tag{20}$$

and

$$G^{\mathrm{R}}_{N+1\,0} = G^{\mathrm{R},(1)}_{N+1\,1}\, H_{10}\, G^{\mathrm{R}}_{00}\ , \tag{21}$$

where $g^{\mathrm{R}}_{\mathrm{L}}$ is the surface Greens function of the left lead.

Each step of the algorithm performs inversions and matrix multiplications with matrices of size M_i. Since the computational complexity of matrix inversion and multiplications scales as ${M_i}^3$, the complexity of the RGF algorithm is $\propto \sum_{i=0}^{N+1} M_i^3$. Thus, it scales linearly with the "length" N, and cubically with the "width" M_i of the system.

While for certain geometries the RGF algorithm cannot compete with more specialized algorithms such as MRGF, it is very versatile and easily adapted to many situations, and is thus our method of choice. In the next section we will discuss matrix reordering techniques that improve the runtime of the RGF algorithm considerably and allow the treatment of arbitrary systems.

Matrix Reordering Strategies for Quantum Transport

Graph-theoretical Approaches to Matrix Reordering

As shown in the previous section, the structure of a Hamiltonian matrix H does influence the runtime of the RGF algorithm. Thus, the runtime of the algorithm can potentially be improved by reordering the matrix with a permutation P,

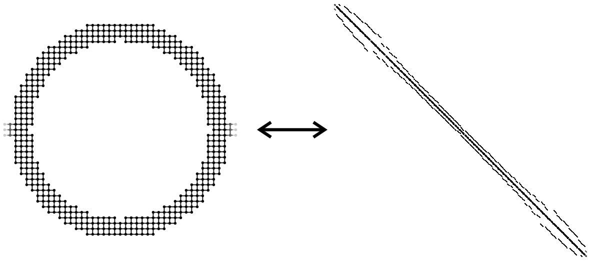
$$H' = P\, H\, P^{-1}\ . \tag{22}$$

At first glance, such an effort may seem pointless: For example, the block-tridiagonal structure naturally associated with a finite difference grid (as discussed in Sect. "Numerical Quantum Transport") leads to a sparse matrix with a small bandwidth, as shown in Fig. 8. However, as we show below, the choice of a suitable permutation P can still lead to a significant speed-up of the RGF algorithm.

For this we define a weight $w(H)$ associated with a given matrix H as

$$w(H) = \sum_{i=0}^{N+1} M_i^3 \quad \text{where } M_i \text{ is the size of block } H_{ii}\ . \tag{23}$$

Optimizing the matrix for the RGF algorithm is then equivalent to minimizing the weight $w(H)$. Since $\sum_{i=0}^{N+1} M_i = N_{\mathrm{grid}}$, where N_{grid} is the total number of grid points, $w(H)$ is minimal, if all M_i are equal, and $M_i = N_{\mathrm{grid}}/(N+2)$. Therefore, a matrix tends to have



Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 8

Example of a sparsity structure of a matrix associated with a finite difference grid. *Black dots* mark non-zero entries. The picture of the finite difference grid can also be interpreted as a graphical representation of a graph

small weight, if the number N of blocks is large, and all blocks are equally sized. The reordering problem of the matrix H is thus summarized as follows:

Matrix reordering problem: Find a reordered matrix H' such, that

1. H'_{00} and H'_{N+1N+1} are blocks given by the left and right leads (as required by the RGF algorithm),
2. H' is block-tridiagonal ($H'_{ij} \neq 0$, iff $j = i+1, i, i-1$),
3. the number N of blocks is as large as possible, and all blocks are equally sized.

These requirements define the optimization problem of reordering the matrix H. Usually, in such optimization problems finding the best solution deterministically is prohibitively expensive, and one has to resort to heuristic strategies.

In order to do that, we reformulate our matrix problem using concepts from graph theory. A *graph* G is a pair $G = (V, E)$, where V is a set of *vertices* i, and E a set of pairs of vertices $(i, j) \in V \times V$. Such a pair is called an *edge*. A graph is called *undirected*, if for every edge $(i, j) \in E$ also $(j, i) \in E$. Two vertices i and j are called *adjacent*, if $(i, j) \in E$. A graph can be visualized by drawing dots for each vertex i and lines connecting these dots for every edge (i,j). It should be noted, that all the pictures of finite difference grids shown so far can directly be interpreted as graphs . There is a natural one-to-one correspondence between graphs and the structure of sparse matrices. For a given $n \times n$ matrix H, we define a graph $G = (V, E)$ with $V = \{1, \ldots, n\}$ and $(i, j) \in E$ iff $H_{ij} \neq 0$. Thus, the symmetric zero–nonzero structure of Hermitian matrices, as considered in quantum transport, leads to associated undirected graphs. An example of such a correspondence between a graph and a matrix has already been shown in Fig. 8. With respect to matrices, graphs are also very convenient for storing and handling sparse matrix data structures on a computer.

In terms of graph theory, matrix reordering corresponds to renumbering the vertices of a graph. Since we are only interested in reordering the matrix in terms of matrix blocks (the order within a block should not matter too much), we define a *partitioning* of G as a set $\{V_i\}$ of disjoint subsets $V_i \subset V$ such that $\bigcup_i V_i = V$ and $V_i \cap V_j = \emptyset$ for $i \neq j$. Using these concepts, we can now reformulate the original matrix reordering problem into a graph partitioning problem:

Graph partitioning problem: Find a partitioning $\{V_0, \ldots, V_{N+1}\}$ of G such that:

1. V_0 and V_{N+1} contain the vertices belonging to left and right leads,
2. there are edges between V_i and V_j iff $j = i+1, i, i-1$ (block-tridiagonality),
3. the number of sets V_i is as large as possible and all sets V_i have the same cardinality. Such a partitioning with all V_i equally sized is called *balanced*.

A partitioning that meets requirement 2 is called a *level set* with levels V_i and appears commonly as an intermediate step in algorithms for bandwidth reduction of a matrix [62,63,64,65]. These algorithms seek to find a level set of minimal width, i. e. $\max_{i=0\ldots N+1} |V_i|$ as small as possible which is equivalent to requirement 3. The major difference between our graph partitioning problem and the bandwidth reduction algorithms is requirement 1: In our case the start and end blocks are given by the geometry of the system, whereas in the bandwidth reduction methods these can be chosen freely. The implications of this difference will be discussed below.

The bandwidth reduction algorithms start with the observation that a *breadth first search* (BFS) starting from any vertex in the graph creates a level set: In our situation, the BFS starts from level V_0 of the left lead. Then, successively for $i = 0, 1, 2, \ldots$, all vertices adjacent to V_i that have not been assigned to a level yet are placed in V_{i+1}. This construction ensures that each level V_i only has edges connecting to vertices in $V_{i+1,i,i-1}$ and thus ensures block-tridiagonality. The search stops, once a vertex adjacent to the right lead is encountered, and all unassigned vertices are placed into the last level V_N. The number of BFS steps determines therefore the maximum number N of blocks and is related to the minimum distance between left and right lead in the graph. However, this construction of the level set also suffers from a serious problem: Depending on the distance between the leads, the last level V_N can potentially contain a large number of vertices leading to a very unbalanced partitioning. In the bandwidth reduction methods, the first and the last vertex are chosen to have (to a good approximation) maximum distance, and thus this problem does not occur there. Hence, conventional bandwidth reduction algorithms can only be applied to quantum transport problems, if the leads are —in terms of the underlying graph —furthest apart.

In this study, we will consider two reordering algorithms: First, the Gibbs–Poole–Stockmeyer (GPS) algorithm [65], a widely used bandwidth reduction algorithm. The GPS algorithm combines the level sets originating from a BFS from both left and right lead to give an optimized level set. Due to the limitations discussed above, it can only be used efficiently for systems with leads far apart. To overcome this difficulty partly, we also propose a second algorithm, later referred to as *BFS partitioning*: The system is bisected recursively by means of a simultaneous BFS from left and right leads. In a bisection process, vertices that are closer (as given by the number of BFS steps) to the left (right) lead are placed into the left (right) level. The resulting two levels are then further bisected recursively until the final level set has been constructed. This algorithm tries to avoid an unbalanced partitioning, as every step tries to create a balanced bisection. We have found this global approach —as opposed to the local approach in the BFS —to yield balanced partitionings for systems where there are only few local minima in the weight $w(H)$. For general systems, a more sophisticated method should be used [66].

Obviously, reordering the matrix will only improve the runtime of the RGF algorithm, if the overhead of the reordering process is small compared to the actual transport calculation. Because of this reason, applying general optimization algorithms to the matrix reordering problem is not an option. Instead, heuristics designed for graph problems give much better performance. The GPS algorithm scales linearly with the number of edges $|E|$, and since in a tight-binding representation $|E| \propto N_{\text{grid}}$, its computational complexity is $\mathcal{O}(N_{\text{grid}})$, whereas the BFS partitioning algorithm scales as $\mathcal{O}(N_{\text{grid}} \log N_{\text{grid}})$. In any case, the scaling is much more favorable than that of the RGF algorithm, $\propto \sum_{i=0}^{N+1} M_i^3$, so that for systems of typical size the overhead of the reordering process becomes negligible, as we will demonstrate in the next section.

Example: Ring Geometry

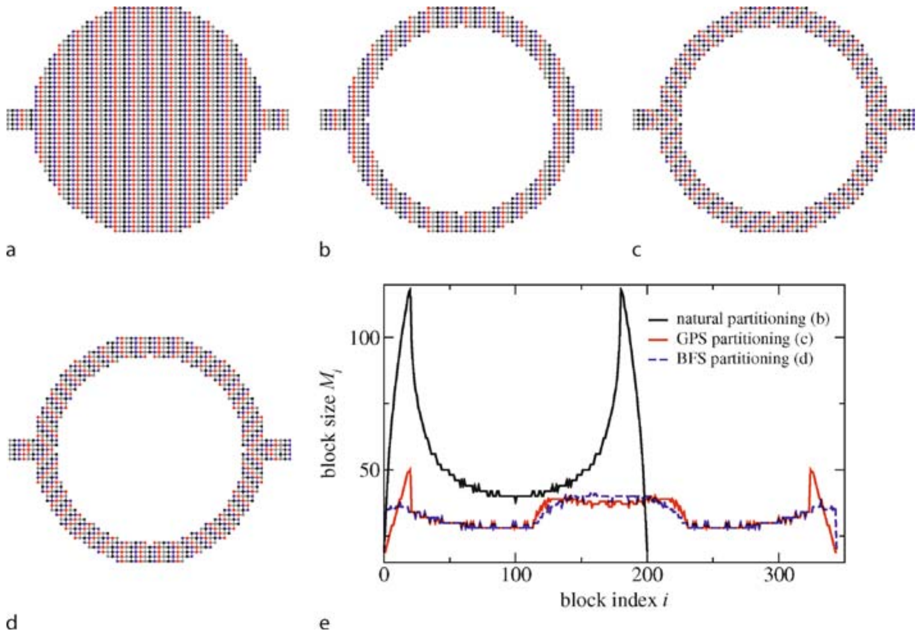
In order to demonstrate the performance of the algorithms discussed above, we consider their application to a ring geometry in finite difference approximation. The performance of the RGF algorithm after matrix reordering is compared with the performance using the ordering that arises naturally in finite difference grids as shown in Fig. 6 (in the remainder referred to as *natural partitioning*).

In Fig. 9a–d we show four different approaches for calculations in a ring geometry. A ring can be treated as a circular cavity (see Fig. 9a), with a large potential on the lattice points inside the inner ring diameter. This approach is easier to implement than a real ring but less efficient, as a large additional number of lattice points enters the calculation. However, this approach has been used frequently in the past, and therefore we also consider its performance. Transport calculations in a real ring require somewhat more bookkeeping because of the non-trivial geometry, but can be easily done describing the grid as a graph. For the circular cavity, we only consider natural partitioning, as shown in Fig. 9a, for the ring we consider natural partitioning (Fig. 9b), GPS partitioning (Fig. 9c) and BFS partitioning (Fig. 9d).

The partitionings in Fig. 9c and d are dramatically different from the natural partitioning. The levels align mainly in vertical or diagonal directions as these are the preferred directions in the square lattice. The number of levels is increased with respect to the natural partitioning, as the distance between both leads is much larger than the number of vertical slices, leading to a larger number of blocks in the block-tridiagonal matrix, as can be seen from Fig. 9e. Both GPS and BFS partitioning lead to a drastically reduced block size with respect to the natural partitioning and the result is rather balanced. Though the actual partitionings in Fig. 9c and d look quite different, with respect to minimizing the weight $w(H)$ they are equally good. The BFS partitioner conforms to the geometric structure, as it puts vertices in levels according to their distance from the leads. Except for a small number of blocks in the beginning and the end of the block-tridiagonal matrix (see Fig. 9e), the GPS partitioner leads to an equally balanced structure, although the partitioning looks quite different. The GPS partitioner works well in this case, as the leads are almost at maximum distance in this ring structure.

We now apply the recursive Greens function algorithm from Sect. "Numerical Quantum Transport" to the different partitionings. In Fig. 10a we show the runtime of the algorithm as a function of the number of lattice points in the leads, which is also the number of lattice points across the arms of the ring. Note that in all cases the runtime includes both the time spent in calculating the matrix reordering and the time spent in the actual transport calculation.

The runtime scales similarly in all cases, as this is the scaling of the RGF algorithm. Nevertheless, the runtimes of the different approaches can differ by a factor that is significant. As expected, the circular cavity is slowest, due



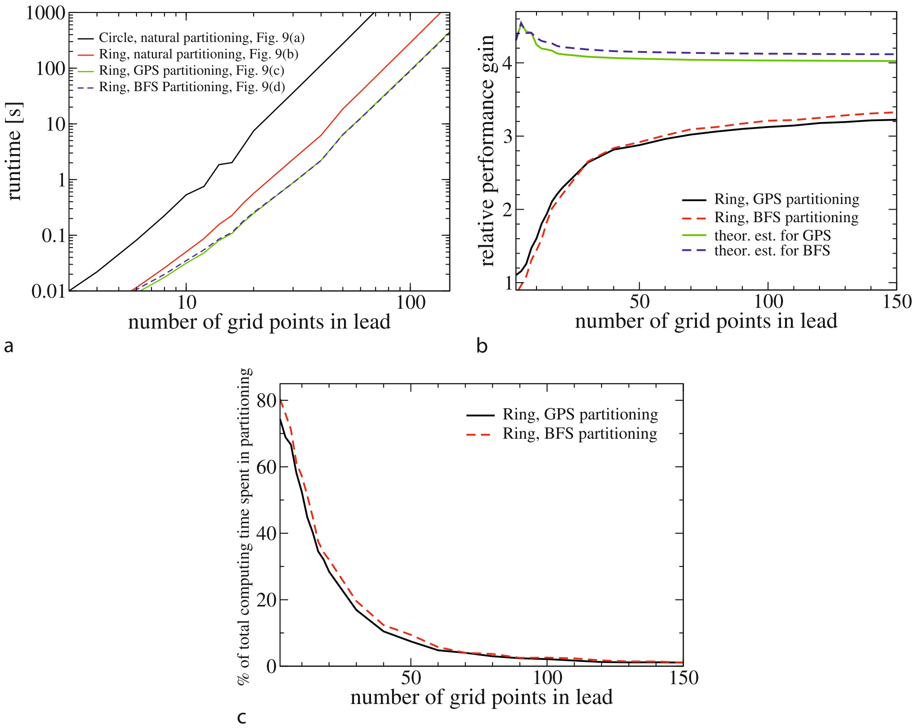
Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 9

Partitioning of the different systems considered in the performance study. a Circular cavity in natural partitioning, where the ring structure is enforced by a potential term in the middle of the cavity, b ring in natural partitioning, c ring in GPS partitioning and d ring in BFS partitioning. In a–d, vertices belonging to the same level in the partitioning are marked with the *same color*, different levels are marked in *alternating colors*. Note that, in order to reveal the partitioning structure more clearly, the grids shown here are much smaller than in a typical calculation. e Size M_i of the blocks H_{ii} in the block-tridiagonal matrix for a ring with 20 lattice points in the leads for natural, GPS and BFS partitioning (the circular cavity is not shown here)

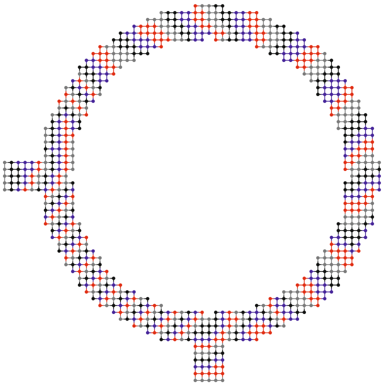
to the extra number of lattice points. GPS and BFS partitionings lead to a rather similar performance that is *significantly* better than the ring in natural partitioning. It outperforms the circular cavity even by a factor of up to 100. In the remainder, we examine the performances of the ring for different partitionings in more detail and leave out the circular cavity. In Fig. 10b we show the relative performance gain of GPS and BFS partitionings over the natural partitioning. Except for the smallest of systems that are too small to be useful in practice, the performance of GPS and BFS partitionings is better than the natural partitionings. Even for moderately sized systems the performance gain through the matrix reordering is approximately 3, with the BFS partitioner being slightly better than the GPS partitioner. In Fig. 10b we also show estimates of the performance gain calculated from the weights $w(H)$ of the different partitionings. These estimates predict a performance gain of approximately 4. For small system sizes, we see deviations from these estimates because of the overhead of the partitioning process, for larger system sizes we almost reach the estimated value. On modern computer architectures, runtime does not only depend on the number of arithmetic operations [67], and thus we do not achieve the full theoretical potential of the reordering, yet still significant improvements. In Fig. 10c we show the fraction of time spent in calculating the matrix reordering with respect to the total computation time, and as expected the overhead becomes negligible already for moderately sized systems. It should be noted that in actual calculations the partitioning is commonly only done once, and transport calculations can be done repeatedly with the same partitioning: Usually one is interested in transport properties depending on some parameters, and these generally do not change the *structure* of the Hamiltonian matrix but only the values of the respective entries. In this case, the partitioning overhead becomes even more irrelevant.

It should be emphasized, that for all the situations, the *same* transport code was used. In addition to the significant speedup through the graph techniques considered here, the abstraction of the system through graph structures allows for very generic transport codes. This is an additional strength of this approach, as the well-established RGF algorithm is thus readily applied to arbitrary systems that would require special treatment otherwise, such as a scattering region with perpendicular leads, as depicted in Fig. 11.

Of course, for any system, an algorithm designed for that special system will probably outperform generic ap-



Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 10
Performance of the partitionings (for rings, Fig. 9): a runtime of the different systems as a function of the number of lattice points in the leads, i. e. as a function of system size. The calculations were performed on a Core2Duo T5500 processor and 1GB of memory, and the runtime includes both partitioning overhead and transport calculation. b Performance gain of GPS and BFS partitionings with respect to natural partitioning and c overhead of the matrix reordering as a function of system size



Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 11
BFS partitioning of a ring with perpendicular leads. Note that in this case the two leads are closer together and thus the number of blocks is reduced. Therefore the block sizes tend to be larger than in the previous examples

proaches. However, the combination of matrix reordering and RGF algorithm can be applied to arbitrary systems and is thus probably the most versatile transport approach. In addition, all the algorithms relying on the block-tridiagonal structure of the Hamiltonian mentioned in Sect. "Numerical Quantum Transport" can benefit from these matrix reordering strategies.

Modern transport calculations tend to be more and more complex and time-consuming. For example, an electronic calculation including the electron spin typically takes $2^3 = 8$ times longer than a calculation on spinless electrons. Furthermore, calculations including disorder involve averages over many disorder configurations. Any increase in computation speed is therefore beneficial, and the added versatility through matrix reordering methods makes these techniques even more useful.

Spin Filtering in Nanostructures

In the introductory example in Sect. "Introduction" it was shown how to realize systems that work as spin switches making use of the interference of wavefunctions propagating clockwise and counterclockwise in Aharonov–Bohm rings with SO-interaction. However, several other device proposals have been put forward utilizing different concepts in order to achieve spin filtering in mesoscopic systems.

A very prominent category is transverse focusing of ballistic electrons/holes in two dimensional electron/hole gases (2DEGs/2DHGs) [68]. In materials that exhibit SO interaction the cyclotron radius of ballistic electrons/holes due to a magnetic field perpendicular to the 2DEG/2DHG depends on the spin state of the charge carriers. Therefore, it is possible to filter out either spin-up or spin-down electrons by an appropriate arrangement of quantum point contacts and a proper choice of the perpendicular magnetic field. Apart from several theoretical treatments of this topic [69,70,71], spin filtering by transverse focusing has already been experimentally verified in a GaAs-based 2DHG [72]. In related experiments, spin-polarized currents in 2DEGs/2DHGs were detected by a setup consisting of point contacts and making use of transverse focusing of the charge carriers [73]. With such a detector it was possible to confirm the presence of spin-polarized currents emitted from mesoscopic quantum dots utilizing quantum interference at high in-plane magnetic fields [74] and from quantum point contacts, which were either made spin sensitive with high in-plane magnetic fields [73] or showed a pronounced "0.7-anomaly" [75].

A further appealing approach to filter spins are three terminal structures, that act as mesoscopic Stern–Gerlach type spin filters [76]. In these devices one of the leads injects spin-unpolarized current and, after passing a region where the spin-degeneracy is lifted, oppositely-polarized output currents exit through the other two leads. This separation of up and down spins can be accomplished, e. g., by utilizing Rashba SO interaction [77,78,79].

However, three terminals are not required to create spin polarized currents. Many devices, as e. g. the AB-ring presented in Sect. "Introduction", typically rely on two terminals only, where transport through tailored geometries with SO interaction [80,81,82], magnetic fields [83,84,85,86] or a combination of both [87] can result in a significant spin filter effect.

As a representative example for the methods mentioned above, in the present section we present spin filtering due to Rashba SO interaction in quantum wires connected to two terminals. We consider a quantum wire in y-direction realized in a 2DEG in the (x, y) plane connected to two nonmagnetic leads. The Hamiltonian of the system, with spatially dependent Rashba SO interaction is given by

$$H_0 = \frac{\hat{p}^2}{2m^*} + \frac{\alpha(x)}{2\hbar}(\hat{\sigma}_x \hat{p}_y - \hat{\sigma}_y \hat{p}_x) + (\hat{\sigma}_x \hat{p}_y - \hat{\sigma}_y \hat{p}_x)\frac{\alpha(x)}{2\hbar} + V(x, y) + U_B(x, y) \, . \tag{24}$$

Here $V(x, y)$ is the lateral transverse confinement potential forming the quantum wire, while $U_B(x, y)$ is an additional electrostatic potential in the system, e. g., realized by gate voltages. Furthermore, $\hat{\sigma}_i$ denote the Pauli spin matrices, and m^* is the effective electron mass of the semiconducting material. We consider a constant Rashba SO interaction strength $(= \alpha_C)$ in the central region of the system, which is connected to two semi-infinite leads on opposite sides, where $\alpha(x)$ is chosen to be zero avoiding ambiguities in the definition of spin current that arise for leads with SO interaction [88]. In order not to introduce additional effects due to an abrupt jump in the SO coupling strength, the parameter $\alpha(x)$ is changed sufficiently smooth from zero to α_C between the leads and the central region. For the numerical calculations presented in the next two sections the Hamiltonian (24) is discretized as shown in Sect. "Numerical Quantum Transport" yielding a tight-binding Hamiltonian on a square grid.

In the rest of this section we investigate the transport properties in the linear response regime due to an infinitesimal bias voltage δU applied between the left and right contact. The charge (C) and spin (S) current in the Landauer–Büttiker formalism are then given by

$$I_{C/S} = G_{C/S} \, T_{C/S} \, \delta U \, ,$$

where $G_C = e^2/h$ and $G_S = e/4\pi$ are the conductance quanta of charge and spin respectively. Since, opposite to charge current, the spin current can be different in the right and left lead [85], here we choose to evaluate the respective currents in the right lead. Then the transmission probabilities $T_{C/S}$ at the Fermi energy are given by

$$\begin{aligned} T_C &= T_{+,+} + T_{+,-} + T_{-,+} + T_{-,-} \, , \\ T_S &= T_{+,+} + T_{+,-} - T_{-,+} - T_{-,-} \, , \end{aligned} \tag{25}$$

where $T_{\sigma,\sigma'} = \sum_{n \in R, n' \in L} |S_{n,\sigma;n',\sigma'}|^2$ is the probability for an electron, injected into the left (L) lead with spin state σ' to be transmitted to the right (R) lead and end up there in spin state σ. In the present and the following section we fix the spin quantization axis to the y-axis. The scattering matrix elements $S_{n,\sigma;n',\sigma'}$ and therefore also the spin resolved transmission probabilities $T_{\sigma,\sigma'}$ are evaluated using the recursive Greens function algorithm presented in Sect. "Numerical Quantum Transport" and "Matrix Reordering Strategies for Quantum Transport".

One general feature of Landauer transport in a quantum wire with SO interaction and non-magnetic leads is the absence of spin-polarized currents in a lead that supports only a single transversal mode. This property can be derived from the invariance of the system under the time-reversal operator $\hat{\mathcal{T}} = -\mathrm{i}\hat{C}\sigma_y$ [89], where $\hat{C}$ is the operator of complex conjugation. For a perfect quantum wire which is translationally invariant in the direction of transport all occupied transversal subbands transmit without reflection and spin polarization is not possible due to SO interaction. However, if backreflection, caused by deviations from a perfect quantum wire, is present, it is possible to observe spin-polarized currents in leads with at least two transversal channels. There the typical mechanism responsible for spin polarization is the mixing of spins from different transversal subbands due to the SO interaction.

In Eq. (24) this translational invariance in y-direction is already broken by the spatially varying SO interaction $\alpha(x)$ even if the quantum wire was perfect in y-direction otherwise, i. e. $V(x, y) = V(y)$ and $U_\mathrm{B}(x, y) = U_\mathrm{B}(y)$. However, if the region where $\alpha(x)$ is turned on/off is sufficiently long, reflection due to the change of $\alpha(x)$ is negligible.

There exist several device proposals relying on this mixing of spins from different transversal subbands due to x-dependent lateral confinement potentials $V(x,y)$ or other superimposed electrostatic potentials $U_\mathrm{B}(x, y)$. These device designs include, e. g., constrictions [80,90], lateral side pockets [81], or electrostatic barriers [82], to name only a few.

In most of those proposals, systems symmetric with respect to inversion of the x-coordinate were considered, i. e. $V(x, y) = V(-x, y)$, $U_\mathrm{B}(x, y) = U_\mathrm{B}(-x, y)$ and $\alpha(x) = \alpha(-x)$. Then the Hamiltonian (24) is left invariant upon application of the symmetry operation

$$\hat{\mathcal{P}} = -\mathrm{i}\,\hat{C}\hat{R}_x\hat{\sigma}_z\ , \tag{26}$$

where the operator $\hat{R}_x$ inverses the x-coordinate. The scattering wavefunctions inside the leads are changed by the operator $\hat{\mathcal{P}}$ in the following way: $\hat{R}_x$ exchanges the leads, i. e., a transversal mode index n is replaced by the corresponding mode index $\bar{n}$ in the other lead. The operator of complex conjugation transforms incoming (outgoing) states into outgoing (incoming) states with complex conjugated amplitude. Moreover, the combined effect of $\hat{C}\sigma_z$ is to rotate a spinor with the coordinates (θ, ϕ) on the Bloch sphere to the coordinates $(\theta, -\phi + \pi)$. Exploiting this invariance of the Hamiltonian one can derive the relation $S_{n,\sigma;n',\sigma'}(E) = S_{\bar{n}',\sigma';\bar{n},\sigma}(E)$ between the elements of the scattering matrix (see [85,89] for related expressions). This results in the equality of the spin flip transmissions $T_{+,-} = T_{-,+}$. Therefore, for those devices to be able to work as a spin filter, in view of Eq. (25) the spin conserving transmissions $T_{+,+}$ and $T_{-,-}$ need to be different.

As an example, in the following we consider Landauer transport through a Rashba SO quantum wire with a constriction. Similar calculations were carried out in [71,80] where it was shown, that this setup is able to produce a spin polarized current of sizeable quantity. It was conjectured, that the mechanism responsible for the spin polarization was the depletion of higher transversal modes of the wire and a subsequent spin dependent repopulation of those modes when traversing the constriction [80]. To experimentally observe the predicted spin polarization the use of a transverse electron focusing technique was suggested [71].

A typical grid (with lattice spacing a) used in the calculation for the symmetric point contact in a wire of width $W = 15a$ is shown in Fig. 12a, where the constriction of length $L_\mathrm{PC} = 10a$ is formed by hard-wall potentials:

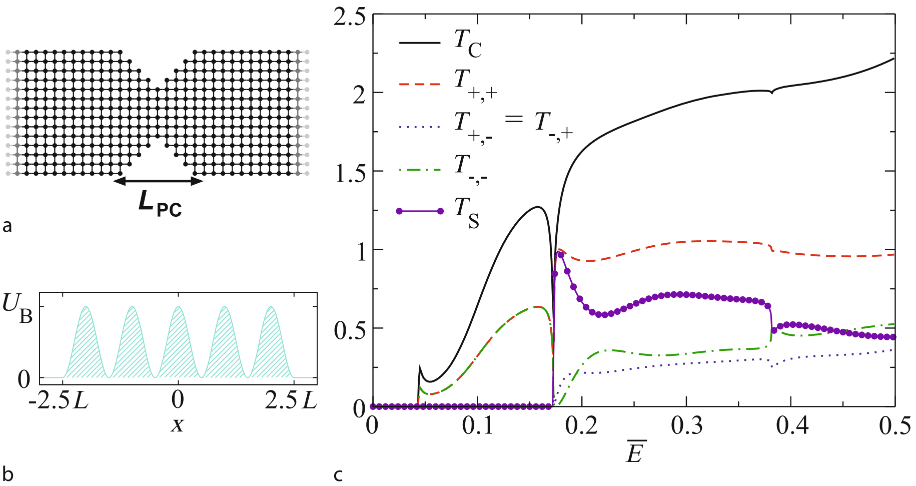
$$V(x, y) = \begin{cases} 0 & \text{for } C(x) < y < 15a - C(x) \\ \infty & \text{else} \end{cases}$$

with

$$C(x) = \begin{cases} 2.05a\left(1 - \cos\left(\frac{2\pi\left(x + \frac{L_\mathrm{PC}}{2}\right)}{L_\mathrm{PC}}\right)\right) & \text{for } |x| < \frac{L_\mathrm{PC}}{2} \\ 0 & \text{otherwise}\,. \end{cases} \tag{27}$$

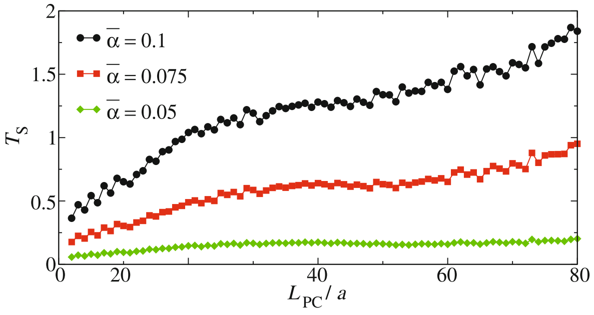
Additionally $U_\mathrm{B}(x, y)$ is set to zero. In Fig. 12c we present the relevant transmission probabilities with respect to the Fermi energy E for this system. There we observe that the total conductance T_C is reduced in comparison with that of a perfect quantum wire. In the latter case T_C exhibits sharp steps due to conductance quantization [91,92] which are washed out here due to tunneling processes through the constriction. Furthermore, for energies, where only a single transversal mode is supported in the quantum wire, the spin transmission vanishes as expected, i. e. $T_{+,+} = T_{-,-}$ for energies $\bar{E} \lesssim 0.175$. Also the relation $T_{+,-} = T_{-,+}$ is fulfilled as required by the symmetry of the setup. Finally, at energies where a new transversal mode opens up ($\bar{E}_2 \approx 0.18$, $\bar{E}_3 \approx 0.39$) dips in the transmission probabilities become apparent. Those dips can be explained by interference of localized states in the central region where $\alpha(x) = \alpha_\mathrm{C}$ and the scattering states in the quantum wire where $\alpha(x) = 0$ [93].

In order to study the influence of the adiabaticity of the constriction on the degree of spin polarization that can be extracted, in Fig. 13 we show T_S as a function of the length of the constriction L_PC for several values of α_C. For all of the curves we clearly observe an increase in spin transmission with increasing L_PC, in accordance with the mechanism suggested in [80]. There it was argued in the context of Landau–Zener transitions that the repopulation of higher transversal subbands will be more efficient for more



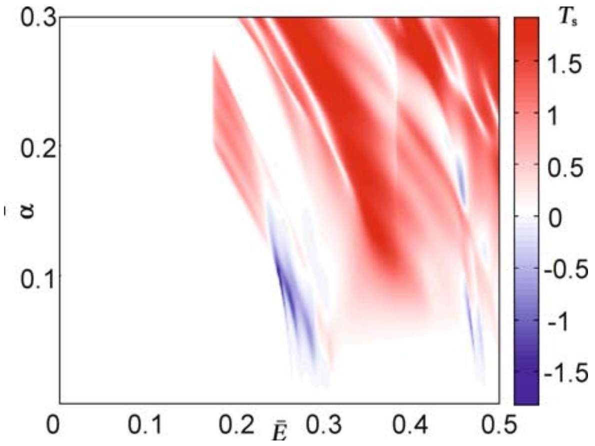
Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 12

Panel a The square lattice discretization of a quantum wire (width $W = 15a$) with a single constriction of length $L_{PC} = 10a$, see also Eq. (27). Panel b Periodic array of $N = 5$ electrostatic barriers with period length L and barrier height U_B. Panel c Charge T_C, spin T_S and spin resolved transmission probabilities $T_{\sigma,\sigma'}$ for the system depicted in panel a) and specified in the text at fixed SO interaction strength $\bar{\alpha} = [(m^* a)/\hbar^2]\,\alpha_C = 0.1$ with respect to the Fermi energy $\bar{E} = [(2m^* a^2)/\hbar^2]\,E$. The nth transversal mode in the wire opens at $\bar{E}_n = (\pi^2 n^2)/(W/a)^2$



Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 13

Spin transmission probability T_S at fixed injection energy $\bar{E} = 0.25$ within the second transversal mode for three different SO interaction strengths $\bar{\alpha} = 0.1$ (*black dots*), $\bar{\alpha} = 0.075$ (*red squares*) and $\bar{\alpha} = 0.05$ (*green diamonds*) with respect to the length of the constriction L_{PC}



Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 14

Spin transmission probability T_S plotted as a function of the Fermi energy $\bar{E}$ and SO interaction strength $\bar{\alpha}$ for a quantum wire with $N = 5$ electrostatic barriers of length $L = 10a$ and height $\bar{U}_{Barr} = [(2m^* a^2)/\hbar^2]\,U_{Barr} = 0.2$

adiabatic constrictions or barriers, resulting in a higher degree of spin polarization. For $\bar{\alpha} = [(m^* a)/\hbar^2]\,\alpha_C = 0.1$ the spin transmission even approaches the highest possible value $T_S = 2$. One drawback of the presented system is its restriction to unidirectional spin polarization. In agreement with the model of [80], Fig. 12 and 13 give evidence that $T_S \geq 0$ for the parameter range considered. This limitation to output current with fixed spin polarization direction restricts the usability of the spin filter to special purposes. A possible way to circumvent this constraint is the use of a periodic array of electrostatic barriers [82], which we now briefly investigate. Figure 14 shows the spin transmission of a straight hard-wall quantum wire ($C(x) = 0$ in Eq. (27)) subject to $N = 5$ electrostatic barriers,

$$U_B(x, y) = U_B(x) = \begin{cases} \frac{U_{Barr}}{2}\left(1 - \cos\left(\frac{2\pi\left(x + \frac{L}{2}\right)}{L}\right)\right) \\ \qquad\qquad \text{for } -N\frac{L}{2} < x < N\frac{L}{2} \\ 0 \quad \text{otherwise}\,, \end{cases}$$

(as shown in Fig. 12b). The spin transmission is plotted as a function of the Fermi energy and the SO interaction strength. Again, one observes $T_S = 0$ for energies below $\bar{E}_2$. Furthermore, different regions in parameter space exhibit opposite sign of T_S, enabling to change the sign of the output polarization, e. g., by tuning α_c via gate voltages [25]. This additional functionality is due to resonant tunneling, which is absent for a quantum wire with only a single constriction or barrier.

Pure Spin Current Generation

In the preceding sections we focused on mesoscopic geometries that exhibit functionalities such as spin filtering or spin switching when applying an external bias between the contacts in the system. In the case of a static dc bias in a two terminal geometry those spin currents arise due to different currents of spin-up and spin-down electrons flowing into the direction of the contact with the lower chemical potential. However, over the last few years a growing number of device proposals has been put forward, that exhibit the interesting feature of pure spin current generation, i. e. spin currents in the absence of net charge transport. This intriguing case appears, when the direction of motion of spin-up electrons is opposite to the direction of motion of spin-down electrons and both currents equal in absolute magnitude.

Among the devices sharing the prospect of creating pure mesoscopic spin currents are systems with more than two terminals realized in 2DEGs exhibiting Rashba [2,3] and/or the Dresselhaus [94] SO coupling. Here a charge current is induced in this multiterminal structure by the application of bias voltages between the different contacts of the system. However, if the voltage of one of the leads is adjusted to make it work as a voltage probe, no charge current passes this lead but a pure spin current can appear owing to the SO coupling present in the system. The basic working principle behind these devices is the so-called mesoscopic spin Hall effect [95], a version of the intrinsic spin Hall effect [27], where the typical system size does not exceed the phase coherence length of the electrons. The SO interaction leads to different transport dynamics for different spin species, which can be used to extract the desired pure spin currents by a clever design of the multiterminal geometry [21,96,97].

Complementary to the generation of pure spin current in multiterminal geometries, there are other types of devices not relying on the application of a net dc-bias. In spin pumping, the cyclic variation of two or more system parameters, such as e. g. gate voltages, induces spin-polarized currents at zero bias, where the induced charge current can be tuned to disappear, leaving pure spin currents. Several realizations of spin pumps in mesoscopic systems have been proposed, relying on SO interaction [98,99] or the Zeeman coupling of electrons to external magnetic fields [100]. The latter proposal has been experimentally confirmed [101] by detecting spin-polarized currents making use of a transverse electron focusing technique [73] mentioned in the previous section.

Complementary to pumps, ratchets only require a single driving parameter to achieve directed transport, and the current direction can be switched upon tuning external parameters such as temperature. In addition to the requirement of a broken spatial symmetry the ratchet has to be operated out of thermal equilibrium. The concept of particle ratchets, which has been addressed in numerous works [102], has recently been extended to systems called spin ratchets. To be specific, the mesoscopic spin ratchets proposed so far [82,85,103], are based on a quantum wire realized in a 2DEG. Between the two contacts attached to the quantum wire an ac bias voltage $U_R(t)$ is applied (rocking ratchet) with zero net (time-averaged) bias. Furthermore, in the central region of the quantum wire the spin degeneracy is lifted due to either SO interaction [82] or the Zeeman coupling to an external magnetic field [85,103]. Upon appropriate choice of the system geometry and tuning of the external driving the charge transported in the forward ($U_R > 0$) and backward bias ($U_R < 0$) situation can be made equal allowing for spin currents in the absence of net charge transport.

In the following we outline the model for the spin ratchets introduced in [82,85,103]. There, driving with a period t_0 is considered. It is implied, that this period is much larger than characteristic time scales related to the electron transport through the quantum wire. Therefore, the system is assumed to be in a steady state at every instance of time, and the Landauer–Büttiker approach to transport is used for the computation of the ratchet currents. To be specific, we consider an unbiased square wave driving $U_R(t) = U_0 \operatorname{sign}[\sin(2\pi t/t_0)]$, where $U_R(t)$ is restricted to the values $\pm U_0$. The net current is then given by the average of the steady-state currents in the two rocking situations (labeled $+U_0$ and $-U_0$ respectively) for both charge and spin,

$$\langle I_{C/S}(U_0)\rangle = \frac{I_{C/S}(+U_0) + I_{C/S}(-U_0)}{2} . \tag{28}$$

Since the spin ratchet effect requires nonlinear transport [103], i. e. finite bias voltages, the Hamiltonian (24) introduced in the previous section has to be extended to additionally include the effective electrostatic potential in the conductor due to the applied bias. Therefore, we add the term $H_R = eU_R g(x, y; U_R)$ to the Hamiltonian (24) yielding the full Hamiltonian at finite bias:

$$H = H_0 + H_R\,. \tag{29}$$

Furthermore at finite bias a generalized version of the expressions for charge and spin current valid at $U_R \neq 0$ has to be used. For coherent Landauer transport those currents can be obtained from an integration of the transmission probabilities over the Fermi window [1]. Finally, the averaged charge $\langle I_C\rangle$ and spin $\langle I_S\rangle$ currents can be written as [85]

$$\langle I_{C/S}(U_0)\rangle = \frac{G_{C/S}}{2e}\int_{E_C}^{\infty} dE\,\Delta f(E, U_0)\Delta T_{C/S}(E, U_0)\,.$$

Here E_C is the energy of the conduction band edge and $\Delta f(E, U_0) = [f(E, E_F + eU_0/2) - f(E, E_F - eU_0/2)]$ is the difference between the Fermi functions of the leads at bias voltage U_0, defining the Fermi window. The averaged charge/spin transmission is just the difference between the steady state transmissions in the two rocking situations:

$$\Delta T_{C/S}(E, U_0) = T_{C/S}(E, +U_0) - T_{C/S}(E, -U_0)\,. \tag{30}$$

Considering the Hamiltonian (29) we now show under what conditions the net charge transported after one full rocking period is zero, i. e., $\langle I_C(U_0)\rangle = 0$. If the electrostatic potentials $V(x, y)$, $U_B(x, y)$ and the Rashba SO strength $\alpha(x)$ are invariant under inversion of the x-coordinate,

$$\begin{aligned} V(x, y) &= V(-x, y)\,,\\ U_B(x, y) &= U_B(-x, y)\,,\\ \alpha(x) &= \alpha(-x)\,, \end{aligned} \tag{31}$$

it is appropriate to assume that the electrostatic potential distribution due to the finite applied voltage $g(x, y; U_R)$ also possesses this symmetry. Then the total Hamiltonian (29) is invariant under the action of the symmetry operation $\hat{\mathcal{P}} = -i\hat{C}\hat{R}_U\hat{R}_x\sigma_z$ where R_U switches the sign of the bias voltage ($\pm U_0 \leftrightarrow \mp U_0$), yielding the relation

$$T_{\sigma,\sigma'}(E, \pm U_0) = T_{\sigma',\sigma}(E, \mp U_0)$$

between the spin-resolved transmission probabilities in the two rocking situations [85]. Inserting this relation into Eq. (30), we observe that the expression for the net charge transmission ΔT_C is zero, resulting in vanishing net charge current. Furthermore it can be used to simplify the expression for the net spin current:

$$\begin{aligned}\langle I_S(U_0)\rangle = &\frac{G_S}{e}\int_{E_C}^{\infty} dE\,\Delta f(E; U_0)\\ &\times [T_{+,-}(E, +U_0) - T_{-,+}(E, +U_0)]\,.\end{aligned}$$

At $U_0 = 0$ the Hamiltonian (29) reduces to Eq. (24), which is invariant under the operation of Eq. (26), yielding $T_{+,-}(E, 0) = T_{-,+}(E, 0)$. This absence of net spin current $\langle I_S(U_0 = 0)\rangle = 0$ in the linear response regime is in agreement with the theoretical prediction [85]. However, for finite rocking voltages $U_0 \neq 0$ the additional potential introduced via H_R breaks this symmetry and therefore enables different spin flip transmissions and thus a resulting net spin current.

We now turn our attention to the quantum wire shown in Fig. 12a. For this system, which exhibits the symmetries of Eq. (31), we perform numerical calculations at finite bias U_0, in order to confirm its operability as a spin ratchet. In general, the function $g(x, y; U_R)$ has to be obtained by self-consistently solving the Schrödinger and Poisson equation of the system. However, in the present treatment we make use of a simple model for $g(x, y; U_R)$ assuming a linear voltage drop in the region where the point contact is formed in the quantum wire:

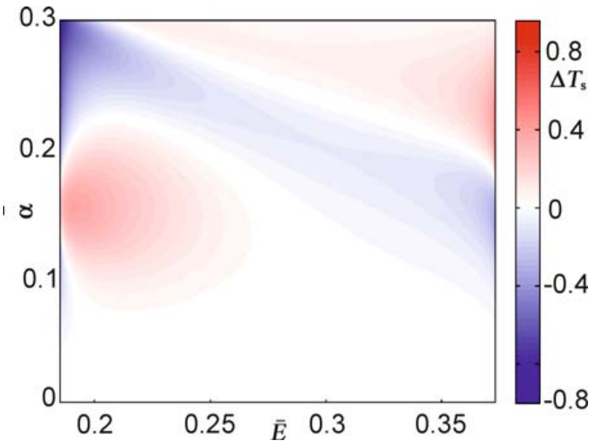
$$g(x, y; U_R) = g(x) = \begin{cases} \frac{1}{2} & \text{for } x < -\frac{L_{PC}}{2}\,,\\ -\frac{x}{L_{PC}} & \text{for } -\frac{L_{PC}}{2} < x < \frac{L_{PC}}{2}\,,\\ -\frac{1}{2} & \text{for } x > \frac{L_{PC}}{2}\,. \end{cases}$$

It is well known that a constriction in a quantum wire acts as an effective potential barrier constituting a region where the voltage applied across the wire is likely to drop. Since the bias voltages we consider are small compared to the energy shift introduced by the constriction, this assumption of a linear voltage drop should be an appropriate approximation for the actual distribution of the electrostatic potential in this wire [104]. Self-consistent calculations have confirmed that this approximation is valid [105].

In Fig. 15 we plot the net spin transmission ΔT_S as a function of $\bar{\alpha}$ and the injection energy in the range, where both leads support two transversal modes. We observe that $|\Delta T_S|$ reaches values of up to 0.9 in the parameter range shown. Furthermore, for a given value of injection energy, the sign of ΔT_S can be switched by changing α_C. Since the strength of the Rashba SO interaction α_C can be tuned via gate voltages [25] the presented system offers also the possibility for experimentally steering and switching the spin current direction.

Future Directions

In the present work we outlined general theoretical and computational concepts of coherent spin-dependent transport at low temperatures and focussed, with regard to numerical examples and possible experimental realiza-



Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 15

Net ratchet spin transmission probability $\Delta T_S(U_0) = 2\left(T_{+,-}(+U_0) - T_{-,+}(+U_0)\right)$ for a stripe with a junction (see text) presented as a function of injection energy $\bar{E}$ and SO interaction strength $\tilde{\alpha}$ at finite applied voltage $eU_0 = 0.02\,\hbar^2/(2m^* a^2)$. Note the sign change in the spin transmission upon tuning the SO coupling

tions, onto ballistic two-dimensional nanostructures based on non-magnetic high-mobility semiconductors.

In order to experimentally achieve high spin polarizations and reasonable spin currents, if possible at room temperature, broad efforts are made to investigate and design novel materials for spintronics. Here, prominent and promising examples, both with respect to fundamental physics and possible applications, are magnetic semiconductors such as GaMnAs [106] or semimagnetic materials with huge g-factors, for instance HgTe [107]. Charge transport in these materials is based on holes. However, relatively few theoretical papers deal with phase coherence effects for hole (spin) transport, though the rich band structure and the interplay between heavy and light hole (or electron- and hole-like) degree of freedoms promise interesting additional features.

The theoretical methods for quantum transport, presented here in the context of mesoscopic systems, are also applied and extended to treat transport in a further prospective field, namely through single-molecule junctions [108], for instance (break) junctions with a molecule bridging the gap between two leads or scanning tunneling microscope measurements of tunnel current through molecules at surfaces. In *Molecular Spintronics* [44,109] spin effects in transport through molecules are addressed. This subfield of spin electronics is still in its infancy. On the computational side these systems pose considerable problems since an adequate approach requires an appropriate description of the electronic and possibly vibrational properties of the molecule including the coupling to and effects of the leads. Whether (spin) DFT calculations for such an embedded molecule, combined with Landauer-type transport calculations, are appropriate, remains to be an issue, in particular if charging or non-equilibrium effects are involved.

As a further future direction we expect that spin transport in graphene, monolayers of graphite, may evolve as another future research line. After its experimental discovery in 2004 [110], graphene has gained much experimental and huge theoretical attention owing to its many exotic properties such as the massless charge carriers, internal spin-like degree of freedoms and unconventional transport characteristics. Also first experiments on graphene-based nanoconductors, e. g., measurements of the Aharonov–Bohm effect in graphene rings [111], are on their way. Graphene is also viewed as a prospective candidate for spin-electronics, since the spin decoherence and spin relaxation times in graphene are expected to be long [112,113]. Recent promising experiments already succeeded in injecting spins from ferromagnetic metallic contacts into graphene, although the conductance mismatch between graphene and the ferromagnetic leads is expected to suppress the efficiency. Recent theoretical proposals predict efficient spin injection into bulk graphene from graphene ribbons employing the occurrence of current-carrying spin-polarized edge states in the ribbons [114].

Acknowledgments

Parts of the results presented are based on previous work together with D. Bercioux, D. Frustaglia, and A. Pfund. We further thank I. Adagideli, M. Grifoni and D. Weiss for useful discussions. Two of us, KR and MW, acknowledge financial support by the Deutsche Forschungsgemeinschaft (SFB 689) and MS through the Studienstiftung des Deutschen Volkes.

Bibliography

Primary Literature

1. Datta S (2002) Electronic Transport in Mesoscopic Systems. Cambridge University Press, Cambridge
2. Rashba EI (1960) Properties of semiconductors with an extremum loop .1. cyclotron and combinational resonance in a magnetic field perpendicular to the plane of the loop. Sov Phys Solid State 2:1109–1122
3. Bychkov YA, Rashba EI (1984) Oscillatory effects and the magnetic susceptibility of carriers in inversion layers. J Phys C: Solid State Phys 17:6039–6045

4. Hentschel M, Schomerus H, Frustaglia D, Richter K (2004) Aharonov–Bohm physics with spin. I. Geometric phases in one-dimensional ballistic rings. Phys Rev B 69:155326
5. Cohen G, Hod O, Rabani E (2007) Constructing spin interference devices from nanometric rings. Phys Rev B 76:235120
6. Loss D, Goldbart P, Balatsky AV (1990) Berry's phase and persistent charge and spin currents in textured mesoscopic rings. Phys Rev Lett 65:1655–1658
7. Stern A (1992) Berry's phase, motive forces, and mesoscopic conductivity. Phys Rev Lett 68:1022–1025
8. Aronov AG, Lyanda-Geller YB (1993) Spin-orbit Berry phase in conducting rings. Phys Rev Lett 70:343–346
9. Qian TZ, Su ZB (1994) Spin-orbit interaction and Aharonov–Anandan phase in mesoscopic rings. Phys Rev Lett 72: 2311–2315
10. Frustaglia D, Hentschel M, Richter K (2001) Quantum transport in nonuniform magnetic fields: Aharonov–Bohm ring as a spin switch. Phys Rev Lett 87:256602
11. Frustaglia D, Hentschel M, Richter K (2004) Aharonov–Bohm physics with spin. II. Spin-flip effects in two-dimensional ballistic systems. Phys Rev B 69:155327
12. Nitta J, Bergsten T (2007) Time reversal Aharonov–Casher effect using Rashba spin-orbit interaction. New J Phys 9:341
13. Morpurgo AF, Heida JP, Klapwijk TM, van Wees BJ, Borghs G (1998) Ensemble-average spectrum of Aharonov–Bohm conductance oscillations: Evidence for spin-orbit-induced Berry's phase. Phys Rev Lett 80:1050–1053
14. Yau JB, De Poortere EP, Shayegan M (2002) Aharonov–Bohm oscillations with spin: Evidence for Berry's phase. Phys Rev Lett 88:146801
15. König M, Tschetschetkin A, Hankiewicz EM, Sinova J, Hock V, Daumer V, Schäfer M, Becker CR, Buhmann H, Molenkamp LW (2006) Direct observation of the Aharonov–Casher phase. Phys Rev Lett 96:076804
16. Grbić B, Leturcq R, Ihn T, Ensslin K, Reuter D, Wieck AD (2007) Aharonov–Bohm oscillations in the presence of strong spin-orbit interactions. Phys Rev Lett 99:176803
17. Frustaglia D, Richter K (2004) Spin interference effects in ring conductors subject to Rashba coupling. Phys Rev B 69:235310
18. Nitta J, Meijer FE, Takayanagi H (1999) Spin-interference device. Appl Phys Lett 75:695–697
19. Mal'shukov AG, Shlyapin VV, Chao KA (1999) Effect of the spin-orbit geometric phase on the spectrum of Aharonov–Bohm oscillations in a semiconductor mesoscopic ring. Phys Rev B 60:R2161–R2164
20. Splettstoesser J, Governale M, Zülicke U (2003) Persistent current in ballistic mesoscopic rings with Rashba spin-orbit coupling. Phys Rev B 68:165341
21. Souma S, Nikolić BK (2005) Spin Hall current driven by quantum interferences in mesoscopic Rashba rings. Phys Rev Lett 94:106602
22. Popp M, Frustaglia D, Richter K (2003) Spin filter effects in mesoscopic ring structures. Nanotechnology 14:347–351
23. Bellucci S, Onorato P (2007) Filtering of spin currents based on a ballistic ring. J Phys: Condens Matter 19:395020
24. Ionicioiu R, D'Amico I (2003) Mesoscopic Stern–Gerlach device to polarize spin currents. Phys Rev B 67:041307
25. Nitta J, Akazaki T, Takayanagi H, Enoki T (1997) Gate control of spin-orbit interaction in an inverted $\mathrm{In}_{0.53}$ $\mathrm{Ga}_{0.47}$ As/$\mathrm{In}_{0.52}$ $\mathrm{Al}_{0.48}$ As heterostructure. Phys Rev Lett 78:1335–1338
26. Schmidt G, Ferrand D, Molenkamp LW, Filip AT, van Wees BJ (2000) Fundamental obstacle for electrical spin injection from a ferromagnetic metal into a diffusive semiconductor. Phys Rev B 62:R4790–R4793
27. Sinova J, Culcer D, Niu Q, Sinitsyn NA, Jungwirth T, MacDonald AH (2004) Universal intrinsic spin Hall effects. Phys Rev Lett 92:126603
28. New J Phys (2007) Focus on Spintronics in Reduced Dimensions. New J Phys 9
29. Corresponding settings, where a precessing magnetization in a ferromagnet emits spin currents, were reviewed by Tserkovnyak Y, Brataas A, Bauer GEW, Halperin BI (2005) Nonlocal magnetization dynamics in ferromagnetic heterostructures. Rev Mod Phys 77:1375
30. Landauer R (1957) Spatial variation of currents and fields due to localized scatterers in metallic conduction. IBM J Res Dev 1:223–231
31. Büttiker M, Imry Y, Landauer R, Pinhas S (1985) Generalized many-channel conductance formula with application to small rings. Phys Rev B 31:6207–6215
32. Stone AD, Szafer A (1988) What is measured when you measure a resistance? – The Landauer forumula revisited. IBM J Res Dev 32:384–413
33. Ferry DK, Goodnick SM (2001) Transport in Nanostructures. Cambridge University Press, Cambridge
34. Fisher DS, Lee PA (1981) Relation between conductivity and transmission matrix. Phys Rev B 23:6851–6854
35. Baranger HU, Stone AD (1989) Electrical linear-response theory in an arbitrary magnetic field: A new fermi-surface formation. Phys Rev B 40:8169–8193
36. Kimball GE, Shortley GH (1934) The numerical solution of Schrödinger's equation. Phys Rev 45:815–820
37. Pauling L, Wilson EB (1935) Introduction to Quantum Mechanics. Dover, New York
38. Havu P, Havu V, Puska MJ, Nieminen RM (2004) Nonequilibrium electron transport in two-dimensional nanostructures modeled using Green's functions and the finite-element method. Phys Rev B 69:115325
39. Bowen RC, Klimeck G, Lake RK, Frensley WR, Moise T (1997) Quantitative simulation of a resonant tunneling diode. J Appl Phys 81:3207–3213
40. Sanvito S, Lambert CJ, Jefferson JH, Bratkovsky AM (1999) General Green's-function formalism for transport calculations with *spd* Hamiltonians and giant magnetoresistance in Co- and Ni-based magnetic multilayers. Phys Rev B 59:11936–11948
41. Luisier M, Schenk A, Fichtner W, Klimeck G (2006) Atomistic simulation of nanowires in the $sp^3d^5s^*$ tight-binding formalism: From boundary conditions to strain calculations. Phys Rev B 74:205323
42. Brandbyge M, Mozos JL, Ordejón P, Taylor J, Stokbro K (2002) Density-functional method for nonequilibrium electron transport. Phys Rev B 65:165401
43. Di Carlo A, Pecchia A, Latessa L, Frauenheim T, Seifert G (2005) Tight-binding DFT for molecular electronics (gDFTB). In: Cuniberti G, Fagas G, Richter K (eds) Introducing Molecular Electronics. Springer, Berlin, pp 153–184
44. Rocha AR, García-Suárez VM, Bailey S, Lambert C, Ferrer J, Sanvito S (2006) Spin and molecular electronics in atomically generated orbital landscapes. Phys Rev B 73:085414

45. Lopez Sancho MP, Lopez Sancho JM, Rubio J (1984) Quick iterative scheme for the calculation of transfer matrices: application to Mo (100). J Phys F: Met Phys 14:1205–s1215
46. Lopez Sancho MP, Lopez Sancho JM, Rubio J (1985) Highly convergent schemes for the calculation of bulk and surface Green functions. J Phys F: Met Phys 15:851–858
47. Krstić PS, Zhang XG, Butler WH (2002) Generalized conductance formula for the multiband tight-binding model. Phys Rev B 66:205319
48. Usuki T, Takatsu M, Kiehl RA, Yokoyama N (1994) Numerical analysis of electron-wave detection by a wedge-shaped point contact. Phys Rev B 50:7615–7625
49. Usuki T, Saito M, Takatsu M, Kiehl RA, Yokoyama N (1995) Numerical analysis of ballistic-electron transport in magnetic fields by using a quantum point contact and a quantum wire. Phys Rev B 52:8244–8255
50. Lambert CJ, Weaire D (1980) Decimation and Anderson localization. Phys Status Solid B 101:591–595
51. Leadbeater M, Lambert CJ (1998) A decimation method for studying transport properties of disordered systems. Ann Phys 7:498–502
52. Mamaluy D, Vasileska D, Sabathil M, Zibold T, Vogl P (2005) Contact block reduction method for ballistic transport and carrier densities of open nanostructures. Phys Rev B 71:245321
53. Thouless DJ, Kirkpatrick S (1981) Conductivity of the disordered linear chain. J Phys C: Solid State Phys 14:235–245
54. Lee PA, Fisher DS (1981) Anderson localization in two dimensions. Phys Rev Lett 47:882–885
55. MacKinnon A (1985) The calculation of transport properties and density of states of disordered solids. Z Phys B 59:385–390
56. Baranger HU, DiVincenzo DP, Jalabert RA, Stone AD (1991) Classical and quantum ballistic-transport anomalies in microjunctions. Phys Rev B 44:10637–10675
57. Lake R, Klimeck G, Bowen RC, Jovanovic D (1997) Single and multiband modeling of quantum electron transport through layered semiconductor devices. J Appl Phys 81:7845–7869
58. Lassl A, Schlagheck P, Richter K (2007) Effects of short-range interactions on transport through quantum point contacts: A numerical approach. Phys Rev B 75:045346
59. Drouvelis P, Schmelcher P, Bastian P (2006) Parallel implementation of the recursive Green's function method. J Comp Phys 215:741–756
60. Rotter S, Tang JZ, Wirtz L, Trost J, Burgdörfer J (2000) Modular recursive Green's function method for ballistic quantum transport. Phys Rev B 62:1950–1960
61. Rotter S, Weingartner B, Rohringer N, Burgdörfer J (2003) Ballistic quantum transport at high energies and high magnetic fields. Phys Rev B 68:165302
62. Cuthill E, McKee J (1969) Reducing the bandwidth of sparse symmetric matrices. In: Proc 24th Nat Conf, ACM, New York, pp 157–172
63. George A (1971) Computer implementation of the finite element method. Tech Rep STAN-CS-71-208, Computer Sci Dept, Stanford Univ, Stanford
64. Liu WH, Sherman AH (1976) Comparative analysis of the Cuthill–McKee and the reverse Cuthill–McKee ordering algorithms for sparse matrices. SIAM J Num Anal 13:198–213
65. Gibbs NE, William G, Poole J, Stockmeyer PK (1976) An algorithm for reducing the bandwidth and profile of a sparse matrix. SIAM J Num Anal 13:236–250
66. Wimmer M, Richter K (2008) Optimal block-tridiagonalization of matrices for coherent charge transport. arXiv:0806.2739v1
67. Whaley RC, Petitet A, Dongarra JJ (2001) Automated empirical optimization of software and the ATLAS project. Parallel Comput 27:3–35
68. van Houten H, Beenakker CWJ, Williamson JG, Broekaart MEI, van Loosdrecht PHM, van Wees BJ, Mooij JE, Foxon CT, Harris JJ (1989) Coherent electron focusing with quantum point contacts in a two-dimensional electron gas. Phys Rev B 39:8556–8575
69. Usaj G, Balseiro CA (2004) Transverse electron focusing in systems with spin-orbit coupling. Phys Rev B 70:041301
70. Govorov AO, Kalameitsev AV, Dulka JP (2004) Spin-dependent transport of electrons in the presence of a smooth lateral potential and spin-orbit interaction. Phys Rev B 70:245310
71. Reynoso A, Usaj G, Balseiro CA (2007) Detection of spin polarized currents in quantum point contacts via transverse electron focusing. Phys Rev B 75:085321
72. Rokhinson LP, Larkina V, Lyanda-Geller YB, Pfeiffer LN, West KW (2004) Spin separation in cyclotron motion. Phys Rev Lett 93:146601
73. Potok RM, Folk JA, Marcus CM, Umansky V (2002) Detecting spin-polarized currents in ballistic nanostructures. Phys Rev Lett 89:266602
74. Folk JA, Potok RM, Marcus CM, Umansky V (2003) A Gate-Controlled Bidirectional Spin Filter Using Quantum Coherence. Science 299:679–682
75. Rokhinson LP, Pfeiffer LN, West KW (2006) Spontaneous spin polarization in quantum point contacts. Phys Rev Lett 96:156602
76. Fabian J, Das Sarma S (2002) Spin transport in inhomogeneous magnetic fields: A proposal for Stern–Gerlach-like experiments with conduction electrons. Phys Rev B 66:024436
77. Kiselev AA, Kim KW (2001) T-shaped ballistic spin filter. Appl Phys Lett 78:775–777
78. Ohe JI, Yamamoto M, Ohtsuki T, Nitta J (2005) Mesoscopic Stern–Gerlach spin filter by nonuniform spin-orbit interaction. Phys Rev B 72:041308
79. Cummings AW, Akis R, Ferry DK (2006) Electron spin filter based on rashba spin-orbit coupling. Appl Phys Lett 89:172115
80. Eto M, Hayashi T, Kurotani Y (2005) Spin polarization at semiconductor point contacts in absence of magnetic field. J Phys Soc Jpn 74:1934
81. Zhai F, Xu HQ (2007) Spin filtering and spin accumulation in an electron stub waveguide with spin-orbit interaction. Phys Rev B 76:035306
82. Scheid M, Pfund A, Bercioux D, Richter K (2007) Coherent spin ratchets: A spin-orbit based quantum ratchet mechanism for spin-polarized currents in ballistic conductors. Phys Rev B 76:195303
83. Song JF, Ochiai Y, Bird JP (2003) Fano resonances in open quantum dots and their application as spin filters. Appl Phys Lett 82:4561–4563
84. Zhai F, Xu HQ (2006) Spin filtering in single magnetic barrier structures revisited. Appl Phys Lett 88:032502

85. Scheid M, Bercioux D, Richter K (2007) Zeeman ratchets: pure spin current generation in mesoscopic conductors with non-uniform magnetic fields. New J Phys 9:401
86. Shi QW, Zhou J, Wu MW (2004) Spin filtering through a double-bend structure. Appl Phys Lett 85:2547–2549
87. Zhai F, Xu HQ (2005) Generation of spin polarization in two-terminal electron waveguides by spin-orbit interaction and magnetic field modulations. Phys Rev B 72:085314
88. Shi J, Zhang P, Xiao D, Niu Q (2006) Proper definition of spin current in spin-orbit coupled systems. Phys Rev Lett 96:076604
89. Zhai F, Xu HQ (2005) Symmetry of spin transport in two-terminal waveguides with a spin-orbital interaction and magnetic field modulations. Phys Rev Lett 94:246601
90. Silvestrov PG, Mishchenko EG (2006) Polarized electric current in semiclassical transport with spin-orbit interaction. Phys Rev B 74:165301
91. van Wees BJ, van Houten H, Beenakker CWJ, Williamson JG, Kouwenhoven LP, van der Marel D, Foxon CT (1988) Quantized conductance of point contacts in a two-dimensional electron gas. Phys Rev Lett 60:848–850
92. Wharam DA, Thornton TJ, Newbury R, Pepper M, Ahmed H, Frost JEF, Hasko DG, Peacock DC, Ritchie DA, Jones GAC (1988) One-dimensional transport and the quantisation of the ballistic resistance. J Phys C: Solid State Phys 21:L209–L214
93. Sánchez D, Serra L (2006) Fano–Rashba effect in a quantum wire. Phys Rev B 74:153313
94. Dresselhaus G (1955) Spin-orbit coupling effects in zinc blende structures. Phys Rev 100:580–586
95. Bardarson JH, Adagideli I, Jacquod P (2007) Mesoscopic spin Hall effect. Phys Rev Lett 98:196601
96. Hankiewicz EM, Molenkamp LW, Jungwirth T, Sinova J (2004) Manifestation of the spin Hall effect through charge-transport in the mesoscopic regime. Phys Rev B 70:241301
97. Sheng L, Sheng DN, Ting CS (2005) Spin-Hall effect in two-dimensional electron systems with Rashba spin-orbit coupling and disorder. Phys Rev Lett 94:016602
98. Governale M, Taddei F, Fazio R (2003) Pumping spin with electrical fields. Phys Rev B 68:155324
99. Sharma P, Brouwer PW (2003) Mesoscopic effects in adiabatic spin pumping. Phys Rev Lett 91:166801
100. Mucciolo ER, Chamon C, Marcus CM (2002) Adiabatic quantum pump of spin-polarized current. Phys Rev Lett 89:146802
101. Watson SK, Potok RM, Marcus CM, Umansky V (2003) Experimental realization of a quantum spin pump. Phys Rev Lett 91:258301
102. Reimann P (2002) Brownian motors: noisy transport far from equilibrium. Phys Rep 361:57–265
103. Scheid M, Wimmer M, Bercioux D, Richter K (2006) Zeeman ratchets for ballistic spin currents. Phys Status Solidi (c) 3:4235
104. McLennan MJ, Lee Y, Datta S (1991) Voltage drop in mesoscopic systems: A numerical study using a quantum kinetic equation. Phys Rev B 43:13846–13884
105. Lassl A, Scheid M, Richter K (2008) Unpublished
106. Ohno H, Shen A, Matsukura F, Oiwa A, Endo A, Katsumoto S, Iye Y (1996) (Ga,Mn)As: A new diluted magnetic semiconductor based on GaAs. Appl Phys Lett 69:363–365
107. König M, Wiedmann S, Brüne C, Roth A, Buhmann H, Molenkamp LW, Qi XL, Zhang SC (2007) Quantum spin Hall insulator state in HgTe quantum wells. Science 318:766–770
108. Cuniberti G, Fagas G, Richter K (eds) (2005) Introducing Molecular Electronics. Springer, Berlin
109. Emberly E, Kirczenow G (2002) Molecular spintronics: spin-dependent electron transport in molecular wires. Chemical Physics 281:311–324
110. Novoselov KS, Geim AK, Morozov SV, Jiang D, Zhang Y, Dubonos SV, Grigorieva IV, Firsov AA (2004) Electric field effect in atomically thin carbon films. Science 306:666–669
111. Russo S, Oostinga JB, Wehenkel D, Heersche HB, Sobhani SS, Vandersypen LMK, Morpurgo AF (2007) Aharonov–Bohm effect in graphene. arXiv:0711.1508v1
112. Huertas-Hernando D, Guinea F, Brataas A (2006) Spin-orbit coupling in curved graphene, fullerenes, nanotubes, and nanotube caps. Phys Rev B 74:155426
113. Min H, Hill JE, Sinitsyn NA, Sahu BR, Kleinman L, MacDonald AH (2006) Intrinsic and Rashba spin-orbit interactions in graphene sheets. Phys Rev B 74:165310
114. Wimmer M, Adagideli I, Berber S, Tománek D, Richter K (2008) Spin transport in rough graphene nanoribbons. Phys Rev Lett 100:177207

Books and Reviews

Bruus H, Flensberg K (2004) Many-body Quantum Theory in Condensed Matter Physics: An Introduction. Oxford University Press, Oxford

Datta S (2002) Electronic Transport in Mesoscopic Systems. Cambridge University Press, Cambridge

Fabian J, Matos-Abiague A, Ertler C, Stano P, Žutić I (2007) Semiconductor Spintronics. Acta Physica Slovaca 57:565–907

Ferry DK, Goodnick SM (2001) Transport in Nanostructures. Cambridge University Press, Cambridge

Stability and Feedback Stabilization

EDUARDO D. SONTAG
Department of Mathematics, Rutgers University,
New Brunswick, USA

Article Outline

Glossary

Stability A globally asymptotically stable equilibrium is a state with the property that all solutions converge to this state, with no large excursions.

Stabilization A system is stabilizable (with respect to a given state) if it is possible to find a feedback law that renders that state a globally asymptotically stable equilibrium.

Lyapunov and control-Lyapunov functions A control-Lyapunov functions is a scalar function which decreases along trajectories, if appropriate control actions are taken. For systems with no controls, one has a Lyapunov function.

Definition of the Subject

The problem of stabilization of equilibria is one of the central issues in control. In addition to its intrinsic interest, it represents a first step towards the solution of more complicated problems, such as the stabilization of periodic orbits or general invariant sets, or the attainment of other control objectives, such as tracking, disturbance rejection, or output feedback, all of which may be interpreted as requiring the stabilization of some quantity (typically, some sort of "error" signal). A very special case, when there are no inputs, is that of stability.

Introduction

This article discusses the problem of stabilization of an equilibrium, which we take without loss of generality to be the origin, for a finite-dimensional system $\dot{x} = f(x, u)$. The objective is to find a *feedback law* $u = k(x)$ which renders the origin of the "closed-loop" system $\dot{x} = f(x, k(x))$ globally asymptotically stable. The problem of stabilization of equilibria is one of the central issues in control. In addition to its intrinsic interest, it represents a first step towards the solution of more complicated problems, such as the stabilization of periodic orbits or general invariant sets, or the attainment of other control objectives, such as tracking, disturbance rejection, or output feedback, all of which may be interpreted as requiring the stabilization of some quantity (typically, some sort of "error" signal). A very special case (when there are no inputs u) is that of stability.

After setting up the basic definitions, we consider *linear* systems. Linear systems are widely used as models for physical processes, and they also play a major role in the general theory of local stabilization. We briefly review pole assignment and linear-quadratic optimization as approaches to obtaining feedback stabilizers.

In general, there is a close connection between the existence of continuous stabilizing feedbacks and smooth *control-Lyapunov functions*, (cfl's), which constitute an extension of the classical concept of Lyapunov functions from dynamical system theory. We discuss the role of clf's in design methods and "universal" formulas for feedback controls.

For nonlinear systems, it has been known since the late 1970s that, in general, there are topological obstructions to the existence of even continuous stabilizers. We review these obstructions, using tools from degree theory.

Finally, we turn to discontinuous stabilization and the associated issue of defining precisely a "solution" for a differential equation with discontinuous right-hand side. We introduce techniques from nonsmooth analysis and differential games, in order to deal with discontinuous controllers. In particular, we discuss the effect of measurement errors on the performance of such controllers.

Preliminaries

In this article, we restrict attention to continuous-time deterministic systems whose states evolve in finite-dimensional Euclidean spaces $\mathbb{R}^n$. (This excludes many other equally important objects of study in control theory: systems which evolve on infinite dimensional spaces and are described by PDE's, systems evolving on manifolds which serve to model state constraints, discrete-time systems described by difference equations, and stochastic systems, among others.) In order to streamline the presentation, we suppose throughout that controls take values in $\mathcal{U} = \mathbb{R}^m$ (constraints in controls would lead to proper subsets $\mathcal{U}$). A *control* (other names: *input*, *forcing function*) is any measurable locally essentially bounded map $u(\cdot)\colon [0, \infty) \to \mathcal{U} = \mathbb{R}^m$. In general, we use the notation $|x|$ for Euclidean norms, and use $\|u\|$ to indicate the essential supremum of a function $u(\cdot)$. For basic terminology and facts about control systems, see [25].

Given a map $f\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ which is locally Lipschitz and satisfies $f(0, 0) = 0$, we consider the associated forced system of ordinary differential equations

$$\dot{x}(t) = f(x(t), u(t)) \,. \tag{1}$$

The maximal solution $x(\cdot)$ of (1) which corresponds to a given initial state $x(0) = x^0$ and to a given control u is defined on some maximal interval $[0, t_{\max}(x^0, u))$, and is denoted by $x(t, x^0, u)$. In the special case when f does not depend on u, we have an unforced system, or system with no inputs

$$\dot{x}(t) = f(x(t)) \,. \tag{2}$$

Unforced systems are associated to a controlled system (1) in two different ways. The first is when one substitutes a feedback law $u = k(x)$ in (1) to obtain a "closed-loop" system $\dot{x} = f(x, k(x))$. The second is when one considers instead the autonomous system $\dot{x} = f(x, 0)$ which models the behavior of (1) in the absence of any controls. All definitions stated for unforced systems are implicitly applied also to systems with inputs (1) by setting $u \equiv 0$; for instance, we define the global asymptotic stability (GAS) property for (2), but we say that (1) is GAS if $\dot{x} = f(x, 0)$ is. For systems with no inputs (2) we write just $x(t, x^0)$ instead of $x(t, x^0, u)$.

Stability and Asymptotic Controllability Stability is one of the most important objectives in control theory, because a great variety of problems can be recast in stability terms. This includes questions of driving a system to a desired configuration (e. g., an inverted pendulum on a cart, to its upwards position), or the problem of tracking a reference signal (such as a pilot's command to an aircraft). We focus in this talk on global asymptotic stabilization.

Recall that the class of $\mathcal{K}_\infty$ functions consists of all $\alpha\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ which are continuous, strictly increasing, unbounded, and satisfy $\alpha(0) = 0$. The class of $\mathcal{KL}$ functions consists of those $\beta\colon \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with the properties that

(1) $\beta(\cdot, t) \in \mathcal{K}_\infty$ for all t, and
(2) $\beta(r, t)$ decreases to zero as $t \to \infty$.

We will also use $\mathcal{N}$ to denote the set of all nondecreasing functions $\sigma\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$. Expressed in terms of such comparison functions, the property of *global asymptotic stability (GAS)* of the origin for a system with no inputs (2) becomes:

$$(\exists \beta \in \mathcal{KL}) \quad |x(t, x^0)| \leq \beta\left(|x^0|, t\right) \quad \forall x^0\,, \forall t \geq 0\,.$$

This definition is equivalent to a more classical "ε-δ" definition usually provided in textbooks, which defines GAS as the combination of stability and global attractivity. For one implication, simply observe that

$$|x(t, x^0)| \leq \beta\left(|x^0|, 0\right)$$

provides the stability (or "small overshoot") property, while

$$|x(t, x^0)| \leq \beta\left(|x^0|, t\right) \underset{t \to \infty}{\longrightarrow} 0$$

gives attractivity. The converse implication is an easy exercise.

More generally, we define what it means for the system with inputs (1) to be (open loop, globally) *asymptotically controllable (AC)* (to the origin). The definition amounts to requiring that for each initial state x^0 there exists some control $u = u_{x^0}(\cdot)$ defined on $[0, \infty)$, such that the corresponding solution $x(t, x^0, u)$ is defined for all $t \geq 0$, and converges to zero as $t \to \infty$, with "small" overshoot. Moreover, we wish to rule out the possibility that $u(t)$ becomes unbounded for x near zero. The precise formulation is as follows.

$$(\exists \beta \in \mathcal{KL})(\exists \sigma \in \mathcal{N}) \quad \forall x^0 \in \mathbb{R}^n \exists u(\cdot)\,,\ \|u\| \leq \sigma\left(\left|x^0\right|\right),$$

$$|x(t, x^0, u)| \leq \beta\left(|x^0|, t\right) \quad \forall\, t \geq 0\,.$$

In particular, (global) asymptotic stability amounts to the existence of $\beta \in \mathcal{KL}$ such that $|x(t, x^0, u)| \leq \beta\left(|x^0|, t\right)$ holds for all $t \geq 0$. A very special case is that of *exponential* stability, in which an estimate of the type $|x(t, x^0, u)| \leq M\mathrm{e}^{-\lambda t}|x^0|$ holds. For linear systems (see below), asymptotic stability and exponential stability coincide. It is a puzzling fact that for general systems, one can find continuous coordinate changes that make asymptotically stable systems exponentially stable [8,13] (a fact of little practical utility, since finding such coordinate changes is as hard as establishing stability to being with).

Feedback Stabilization A map $k\colon \mathbb{R}^n \to \mathcal{U}$ is a *feedback stabilizer* for the system with inputs (1) if k is locally bounded (that is, k is bounded on each bounded subset of $\mathbb{R}$), $k(0) = 0$, and the closed-loop system

$$\dot{x} = f(x, k(x)) \tag{3}$$

is GAS, i. e. there is some $\beta \in \mathcal{KL}$ so that $|x(t)| \leq \beta\left(|x(0)|\,, t\right)$ for all solutions and all $t \geq 0$. (A technical difficulty with this definition lies the possible lack of solutions of (1) when k is not regular enough. We ignore this for now, but will most definitely return to this issue later.)

For example, if (1) is a model of an undamped spring/mass system, where u represents the net effect of external forces, one obvious way to asymptotically stabilize the system is to introduce damping. In control-theoretic terms, this means that we choose $u(t) = k(x(t))$ to be a negative linear function of the velocity. Physically, one may implement a feedback controller by means of an analog device. In the example of the spring/mass system, one could achieve this by adding friction or connecting a dashpot. Alternatively, in modern control technology, one uses a digital computer to measure the state x and compute the appropriate control action to be applied. (There are many implementation issues which arise in digital control and are ignored in our theoretical formulation "$u = k(x)$", among them the effect of delays in the actual computation of the control $u(t)$ to be applied at time t, and the effect of quantization due to the finite precision of measuring devices and the digital nature of the computer. These issues are addressed in the literature, although a comprehensive theoretical framework is still lacking.)

Observe that, obviously, if there exists a feedback stabilizer for (1), then (1) is also AC (we just use $u(t) := k(x(t, x^0))$ as u_{x^0}). Thus, it is very natural to ask whether the converse holds: *is every asymptotically controllable system also feedback stabilizable?*

Linear Systems

A *linear system* is a system (1) for which the map f is linear. In other words, there are two linear transformations $A\colon \mathbb{R}^n \to \mathbb{R}^n$ and $B\colon \mathbb{R}^m \to \mathbb{R}^n$ so that the equations take the form

$$\dot{x} = Ax + Bu\,. \tag{4}$$

Such a system is completely specified once that we are given A and B, which we identify by abuse of notation with their respective matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ with respect to the canonical bases in $\mathbb{R}^n$ and $\mathbb{R}^m$. We also say "the system (A, B)" when referring to (4).

It is natural to look specifically for *linear* feedbacks $k\colon \mathbb{R}^n \to \mathbb{R}^m$ which stabilize a linear system (just as a linear term, inversely proportional to velocity, stabilizes an undamped harmonic oscillator). (In fact, this is no loss of generality, since it can be easily proved for linear systems [25] that if a feedback stabilizer $u = k(x)$ exists, then there also exists a linear feedback stabilizer.) We write $u = Fx$, when expressing $k(x) = Fx$ in matrix terms with respect to the canonical bases. Substituting this control law into (4) results in the equation $\dot{x} = (A + BF)x$. Thus, the mathematical problem reduces to:

given $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$, find $F \in \mathbb{R}^{m\times n}$ such that $A + BF$ is Hurwitz.

(Recall that a Hurwitz matrix is one all whose eigenvalues have negative real parts, and that the origin of the system $\dot{x} = Hx$ is globally asymptotically stable if and only if H is a Hurwitz matrix.) The fundamental stabilization result for linear systems is as follows [25]:

Theorem 1 *A linear system is asymptotically controllable if and only if it admits a linear feedback stabilizer.*

A Remark on Linearization

If the dynamics map f in (1) is continuously differentiable, we may expand to first order $f(x, u) = Ax + Bu + o(x, u)$. Let us suppose that the linearized system (A, B) is AC, and pick a linear feedback stabilizer $u = Fx$, whose existence is guaranteed by Theorem 1. Then, the same feedback law $k(x) = Fx$, when fed back into the original system (1), results in $\dot{x} = f(x, Fx) = (A + BF)x + o(x)$. Thus, k locally stabilizes the origin for the nonlinear system. Of course, the assumption that the linearization is AC is not always satisfied. Systems in which inputs enter multiplicatively, such as those controlling reaction rates in chemical problems, lead to degenerate linearizations. In addition, even if the linearized system (A, B) is AC, in general a linear stabilizer $u = Fx$ will not work as a global stabilizer. For example, the system $\dot{x} = x + x^2 + u$ can be locally stabilized with $u := -2x$, but any linear feedback $u = -fx$ $(f > 1)$ results in an additional equilibrium away from the origin (at $x = f - 1$). Nonlinear feedback must be used (obviously, in this, example, $u = -2x - x^2$ works for global stabilization).

Returning to linear systems, let us note that Theorem 1 is of great interest because (1) there is a simple algebraic test to check the AC property, and (2) there are several practically useful algorithms for obtaining a stabilizing F, including pole placement and optimization, which we sketch next.

Pole Placement

The first technique for stabilization is purely algebraic. In order to simplify this exposition, we will suppose that the system (4) is not just AC but is in fact *controllable*, meaning that every state can be steered, in finite time, to every other state. (Any AC system (4) can be decomposed into two components, of which one is already GAS and the other one is controllable, cf. [25], so this represents no loss of generality.) Controllability is characterized by the property – generically satisfied for pairs (A, B) – that

$$\operatorname{rank}\left[B\middle|AB\middle|A^2B\middle|\dots\middle|A^{n-1}B\right] = n$$

(note that the composite matrix shown has n rows and nm columns).

Two pairs (A, B) and $(\widetilde{A}, \widetilde{B})$ are said to be *feedback equivalent* if there exist $T \in GL(n, \mathbb{R})$, $F_0 \in \mathbb{R}^{m\times n}$, and $V \in GL(m, \mathbb{R})$ so that

$$(A + BF_0)T = T\widetilde{A} \quad \text{and} \quad BV = T\widetilde{B}\,. \tag{5}$$

Feedback equivalence corresponds to changes of basis in the state and control-value spaces (invertible matrices T and V, respectively) and feedback transformations $u = F_0x + u'$, where u' is a new control. (An equivalent way to describe feedback equivalence is by the requirement that two pairs should be in the same orbit under the action of a "feedback group" which is obtained as a suitable semidirect product of $GL(n,\mathbb{R})$, $(\mathbb{R}^{m\times n},+)$, and $GL(m,\mathbb{R})$.) Controllability is preserved under feedback equivalence. Moreover, if (5) holds and if one finds a matrix $\widetilde{F}$ so that $\widetilde{A}+\widetilde{B}\widetilde{F}$ is Hurwitz, then

$$A+BF = T(\widetilde{A}+\widetilde{B}\widetilde{F})T^{-1}$$

is also Hurwitz, where $F := F_0 + V\widetilde{F}T^{-1}$. Thus, the task of finding a stabilizing feedback F can be reduced to the same problem for any pair $(\widetilde{A},\widetilde{B})$ which is feedback equivalent to the given pair (A,B).

One then proceeds to show that there always exists an equivalent pair $(\widetilde{A},\widetilde{B})$ which has a form simple enough that the existence of $\widetilde{F}$ is trivial to establish. In order to find such a pair, it is useful to study the classification of controllable pairs under feedback equivalence. This classification is closely related to Kronecker's theory of "matrix pencils" applied to polynomial matrices $[\lambda I - A, B] = \lambda[I,0] + [-A,B]$ modulo matrix equivalence, cf. [25]. The orbits under feedback equivalence are in one-to-one correspondence with the possible partitions of $n = \kappa_1 + \ldots + \kappa_r$ into the sum of r positive integers, $r \le m$, and in each orbit one can find a pair $(\widetilde{A},\widetilde{B})$ which is in "controller canonical form", for which F can be trivially found. For simplicity, let is just discuss here the very special case of single-input systems ($m = 1$). For this case, the action of the feedback group is transitive, and each controllable system is feedback equivalent to the following special system:

$$A := \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \\ 0 & 0 & 0 & \ldots & 0 \end{pmatrix} \quad B := \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} .$$

For this system, a stabilizing feedback is trivial to obtain. Indeed, take the polynomial $p(\lambda) = (\lambda+1)^n = \lambda^n - \alpha_n\lambda^{n-1} - \ldots - \alpha_2\lambda - \alpha_1$. With $F = (\alpha_1,\alpha_2,\alpha_3,\ldots,\alpha_n)$,

$$A+BF := \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \\ \alpha_1 & \alpha_2 & \alpha_3 & \ldots & \alpha_n \end{pmatrix}$$

has characteristic polynomial $(\lambda+1)^n$, and hence is a Hurwitz matrix, as required for stabilization.

Observe that, instead of the particular $p(\lambda)$ which we used, we could have picked any polynomial all whose roots have negative real parts, and the same argument applies. The conclusion is that, not only can we make $A+BF$ Hurwitz, but we can assign to it any desired set of n eigenvalues (as long as they form a set closed under conjugation). This is the reason that the technique is called *eigenvalue placement* (or "pole placement" because the eigenvalues of A are the poles of the resolvent $(\lambda I - A)^{-1}$). See Chap. 5 in [25] for a detailed treatment of the pole placement problem.

Variational Approach

A second technique for stabilization is based on optimal control techniques. We first pick any two symmetric positive definite matrices $R \in \mathbb{R}^{m\times m}$ and $Q \in \mathbb{R}^{n\times n}$ (for instance the identity matrices of the respective sizes). Next, we consider the problem of minimizing, for each initial state x^0 at time $t = 0$, the infinite-horizon cost

$$\mathcal{J}_{x^0}(u) := \int_0^\infty \left\{u(t)'Ru(t) + x(t)'Qx(t)\right\} \mathrm{d}t$$

over all controls $u\colon [0,\infty) \to \mathbb{R}^m$ which make $\mathcal{J}_{x^0}(u) < \infty$, where $x(t) = x(t,x^0,u)$ and prime indicates transpose. The main result from linear-quadratic optimal control (cf. Sect. 8.4 in [25]) implies that, for AC systems, there is a unique solution u to this problem, which is given in the following form: there is a matrix $F \in \mathbb{R}^{m\times n}$ such that solving $\dot{x} = (A+BF)x$ with $x(0) = x^0$ gives that $u(t) := Fx(t)$ minimizes $\mathcal{J}_{x^0}(\cdot)$. Moreover, this F stabilizes the system (which is intuitively to be expected, since $\mathcal{J}_{x^0}(u) < \infty$ implies that solutions $x(t)$ must be in L^2), and F can be computed by the formula

$$F := -R^{-1}B'P\,, \tag{6}$$

where P is a symmetric and positive definite solution of the *Matrix Algebraic Riccati Equation*

$$PBR^{-1}B'P - PA - A'P - Q = 0\,. \tag{7}$$

A Sufficient Nonlinear Condition

Although of limited applicability, it is worth remarking that there is a partial extension to nonlinear systems of the stabilization method which was just described. For simplicity, we specialize our discussion to control-affine systems, i. e., those for which the input appears only in an affine form. This class is sufficient for the study of most forced mechanical systems. The Eq. (1) becomes:

$$\dot{x} = g_0(x) + \sum_{i=1}^m u_i g_i(x) = g_0(x) + G(x)u\,.$$

(See Sect. 8.5 in [25] for general f, and for proofs). We now pick two continuous functions $Q\colon \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ and $R\colon \mathbb{R}^n \to \mathbb{R}^{n\times n}$, so that $R(x)$ is a symmetric positive definite matrix for each x.

In general, we say that a continuous function

$$V\colon \mathbb{R}^n \to \mathbb{R}_{\geq 0}$$

is *positive definite* if $V(x) = 0$ only if $x = 0$, and it is *proper* (or "weakly coercive") if for each $a \geq 0$ the set $\{x | V(x) \leq a\}$ is compact, or, equivalently, $V(x) \to \infty$ as $|x| \to \infty$ (radial unboundedness). Given any such V which is also differentiable, we denote the vector function whose components are the directional derivatives of V in the directions of the various control vector fields g_i, $i \geq 1$ by:

$$L_G V(x) := \nabla V(x) G(x) = \left(L_{g_1} V(x), \ldots, L_{g_m} V(x)\right) ,$$

and also write $L_{g_0} V(x) := \nabla V(x) g_0(x)$.

We consider the following PDE on such functions V:

$$\forall x \quad Q(x) + L_{g_0} V(x) - \frac{1}{4} L_G V(x) R(x)^{-1} \left(L_G V(x)\right)' = 0 . \tag{8}$$

This reduces to the Algebraic Riccati Eq. (7) in the special case of linear systems, quadratic $V(x) = x'Px$ and $Q(x) = x'Qx$, and constant matrices $R(x) \equiv R$. We also take the following generalization of the feedback law (6):

$$k(x) := -\frac{1}{2} R(x)^{-1} \left(L_G V(x)\right)' . \tag{9}$$

Finally, we assume that Q is a positive definite function. One then has [25]:

Theorem 2 *Suppose that V is a twice continuously differentiable, positive definite, and proper solution of the PDE (8). Then, k defined by (9) stabilizes the system.*

This theorem arises from the following optimization problem: for each state $x^0 \in \mathbb{R}^n$, minimize the cost

$$\mathcal{J}_{x^0}(u) := \int_0^\infty u(t)' R(x(t)) u(t) + Q(x(t)) \mathrm{d}t ,$$

where $x(t) = x(t, x^0, u)$, over all those controls $u\colon [0, \infty) \to \mathcal{U}$ for which the solution $x(t, x^0, u)$ of (1) is defined for all $t \geq 0$ and satisfies $\lim_{t\to\infty} x(t) = 0$. Under the above assumptions, and as for linear systems, one also concludes that for each state x^0 the solution of $\dot{x} = f(x, k(x))$ with initial state $x(0) = x^0$ exists for all $t \geq 0$, the control $u(t) = k(x(t))$ is optimal, and $V(x^0)$ is the optimal cost from initial state x^0. Moreover, the formula for k arises from the Hamilton–Jacobi–Bellman equation of optimal control theory, because

$$k(x) = \underset{u}{\operatorname{argmin}} \left\{\nabla V(x) \cdot f(x, u) + u' R(x) u + Q(x)\right\}$$

when $f(x, u) = g_0(x) + G(x)u$.

There are applications where this method has proven useful. Unfortunately, however, and in contrast to the linear case, in general there exists no positive definite, proper, and C^2 solution V of the above PDE. On the other hand, the formula (9) does appear, with variations, in other contexts, including generalizations of the idea of adding damping to systems, cf. Sect. 5.9 in [25], and, more generally, the use of auxiliary positive definite and proper functions V, in similar roles, will be central to the control-Lyapunov ideas discussed later.

Nonlinear Systems: Continuous Feedback

One of the central topics which we will address here concerns possibly discontinuous feedback laws k. Before turning to that subject, however, we study continuous feedback. When dealing with linear systems, linear feedback is natural, and indeed sufficient from a theoretical standpoint, as shown by the results just reviewed. However, for our general study, major technical questions arise in even deciding on just what degree of regularity should be imposed on the feedback maps k.

It turns out that the precise requirements away from 0, say asking whether k is merely continuous or smooth, are not very critical; it is often the case that one can "smooth out" a continuous feedback (or, even, make it real-analytic, via Grauert's Theorem) away from the origin. So, in order to avoid unnecessary complications in exposition due to nonuniqueness, let us call a feedback k *regular* if it is locally Lipschitz on $\mathbb{R}^n \setminus \{0\}$. For such k, solutions of initial value problems $\dot{x} = f(x, k(x))$, $x(0) = x^0$, are well defined (at least for small time intervals $[0, \varepsilon)$) and, provided k is a stabilizing feedback, are unique (cf. Exercise 5.9.9 in [25]).

On the other hand, behavior at the origin cannot be "smoothed out" and, at zero, the precise degree of smoothness plays a central role in the theory [12]. For instance, consider the system

$$\dot{x} = x + u^3 .$$

The continuous (and, in fact, smooth away from zero) feedback $u = k(x) := -\sqrt[3]{2x}$ globally stabilizes the system (the closed-loop system becomes $\dot{x} = -x$). However, there is no possible stabilizing feedback which is differentiable at the origin, since $u = k(x) = O(x)$ implies that

$$\dot{x} = x + O\left(x^3\right)$$

about $x = 0$, which means that the solution starting at any positive and small point moves to the right, instead of towards the origin. (A general result, assuming that A has no purely imaginary eigenvalues, cf. [25], Section 5.8, is that if – and only if – $\dot{x} = Ax + Bu + o(x, u)$ can be locally asymptotically stabilized using a feedback which is differentiable at the origin, the linearization $\dot{x} = Ax + Bu$ must be AC itself. In the example that we gave, this linearization is just $\dot{x} = x$, which is not AC.)

We now turn to the question of existence of regular feedback stabilizers. We first study a comparatively trivial case, namely systems with one state variable and one input. After that, we turn to multidimensional systems.

The Special Case $n = m = 1$

There are algebraic obstructions to the stabilization of $\dot{x} = f(x, u)$ if the input u appears nonlinearly in f. Ignoring the requirement that there be a $\sigma \in \mathcal{N}$ so that controls can be picked with $\|u\| \leq \sigma(|x^0|)$, asymptotic controllability is, for $n = m = 1$, equivalent to:

$$(\forall x \neq 0)(\exists u)\ xf(x, u) < 0 \tag{10}$$

(this is proved in [28]; it is fairly obvious, but some care must be taken to deal with the fact that one is allowing arbitrary measurable controls; the argument proceeds by first approximating such controls by piecewise constant ones). Let us introduce the following set:

$$\mathcal{O} := \{(x, u) | xf(x, u) < 0\} \ ,$$

and let $\pi\colon (x, u) \mapsto x$ be the projection into the first coordinate in $\mathbb{R}$. Then, (10) is equivalent to:

$$\pi\mathcal{O} = \mathbb{R} \setminus \{0\} \ .$$

(One can easily include the requirement "$\|u\| \leq \sigma(|x^0|)$" by asking that for each interval $[-K, K] \subset \mathbb{R}$ there must be some compact set $C_K \subset \mathbb{R}^2$ so that $[-K, K] \subseteq \pi(C_K)$. For simplicity, we ignore this technicality.) In these terms, a stabilizing feedback is nothing else than a locally bounded map $k\colon \mathbb{R} \to \mathbb{R}$ such that $k(0) = 0$ and so that k is a section of π on $\mathbb{R} \setminus \{0\}$:

$$(x, k(x)) \in \mathcal{O} \ \ \forall x \neq 0 \ .$$

For a regular feedback, we ask that k be locally Lipschitz on $\mathbb{R} \setminus \{0\}$.

Clearly, there is no reason for Lipschitz, or for that matter, just continuous, sections of π to exist. As an illustration, take the system

$$\dot{x} = x\left[(u - 1)^2 - (x - 1)\right]\left[(u + 1)^2 + (x - 2)\right] \ .$$

Let

$$\mathcal{O}_1 = \left\{(u + 1)^2 < (2 - x)\right\} \text{ and } \mathcal{O}_2 = \{(u-1)^2 < (x-1)\}$$

(these are the interiors of two disjoint parabolas). Here, $\mathcal{O}$ has three connected components, namely $\mathcal{O}_2$, $\mathcal{O}_1$ intersected with $x < 0$, and $\mathcal{O}_1$ intersected with $x < 0$. It is clear that, even though $\pi\mathcal{O} = \mathbb{R}$, there is no continuous curve (graph of $u = k(x)$) which is always in $\mathcal{O}$ and projects onto $\mathbb{R} \setminus \{0\}$. On the other hand, there exist many possible feedback stabilizers provided that we allow one discontinuity. It is also possible to provide examples, even with $f(x, u)$ smooth, for which an infinite number of switches are needed in any possible stabilizing feedback law. Finally, it may even be possible to stabilize semiglobally with a regular feedback, meaning that for each compact subset K of the state-space there is a regular, even smooth, feedback law $u = k_K(x)$ such that all states in K get driven asymptotically to the origin, but yet it may be impossible to find a single $u = k(x)$ which works globally. See [26] for details.

When feedback laws are required to be continuous at the origin, new obstructions arise. The case of systems with $n = m = 1$ is also a good way to introduce this subject. The first observation is that stabilization about the origin (even if just local) means that we must have, near zero:

$$f(x, k(x)) \begin{cases} > 0 & \text{if} \quad x < 0 \\ < 0 & \text{if} \quad x > 0 \\ = 0 & \text{if} \quad x = 0 \end{cases} .$$

In fact, all that we need is that $f(x_1, k(x_1)) < 0$ for some $x_1 > 0$ and $f(x_2, k(x_2)) > 0$ for some $x_2 < 0$. This guarantees, via the intermediate-value theorem that, if k is continuous, the projection

$$(-\varepsilon, \varepsilon) \to \mathbb{R} \ , \quad x \mapsto f(x, k(x))$$

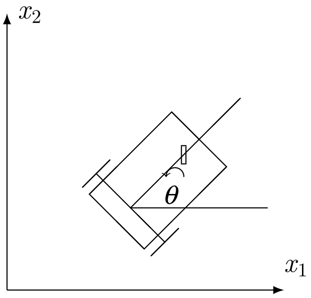
is onto a neighborhood of zero, for each $\varepsilon > 0$. It follows, in particular, that the image of

$$(-\varepsilon, \varepsilon) \times (-\varepsilon, \varepsilon) \to \mathbb{R} \ , \quad (x, u) \mapsto f(x, u)$$

also contains a neighborhood of zero, for any $\varepsilon > 0$ (that is, the map $(x, u) \mapsto f(x, u)$ is open at zero). This last property is intrinsic, being stated in terms of the original data $f(x, u)$ and not depending upon the feedback k. Brockett's condition, to be described next, is a far-reaching generalization of this argument; in its proof, degree theory replaces the use of the intermediate value theorem.

Obstructions and Necessary Degree Conditions

If there are "obstacles" in the state-space, or more precisely if the state-space is a proper subset of $\mathbb{R}^n$, discontinuities in feedback laws cannot in general be avoided,



Stability and Feedback Stabilization, Figure 1
Shopping cart

since the domain of attraction of an asymptotically stable vector field must be diffeomorphic to Euclidean space. But even if states evolve in Euclidean spaces, similar obstructions may arise. These are due not to the topology of the state space, but to "virtual obstacles" implicit in the form of the system equations. These obstacles occur when it is impossible to move instantaneously in certain directions, even if it is possible to move eventually in every direction, the phenomenon of "nonholonomy". As an illustration, let us consider a model for the "shopping cart" shown in Fig. 1 ("knife-edge" or "unicycle" are other names for this example). The state is given by the orientation θ, together with the coordinates x_1, x_2 of the midpoint between the back wheels. The front wheel is a castor, free to rotate. There is a non-slipping constraint on movement: the velocity $(\dot{x}_1, \dot{x}_2)'$ must be parallel to the vector $(\cos\theta, \sin\theta)'$. This leads to the following equations:

$$\begin{aligned}\dot{x}_1 &= u_1 \cos\theta\\ \dot{x}_2 &= u_1 \sin\theta\\ \dot{\theta} &= u_2\end{aligned}$$

where we may view u_1 as a "drive" command and u_2 as a steering control. (In practice, one would implement these controls by means of differential forces on the two back corners of the cart.) The feedback transformation $z_1 := \theta$, $z_2 := x_1 \cos\theta + x_2 \sin\theta$, $z_3 := x_1 \sin\theta - x_2 \cos\theta$, $v_1 := u_2$, and $v_2 := u_1 - u_2 z_3$ brings the system into the system with equations $\dot{z}_1 = v_1, \dot{z}_2 = v_2, \dot{z}_3 = z_2 v_1$ known as "Brockett's example" or "nonholonomic integrator" (yet another change can bring the third equation into the form $\dot{z}_3 = z_1 v_2 - z_2 v_1$). We view the system as having state space $\mathbb{R}^3$. Although a physically more accurate state space would be the manifold $\mathbb{R}^2 \times \mathbb{S}^1$, the necessary condition to be given is of a local nature, so the global structure is unimportant.

This system is (obviously) completely controllable (formally, controllability can be checked using the Lie algebra rank condition, as in e. g. [25], Exercise 4.3.16), and in particular is AC. However, we may expect that discontinuities are unavoidable due to the non-slip constraint, which does not allow moving from, for example the position $x_1 = 0$, $\theta = 0$, $x_2 = 1$ in a straight line towards the origin. Indeed, one then has [3]:

Theorem 3 *If there is a stabilizing feedback which is regular and continuous at zero, then the map $(x, u) \mapsto f(x, u)$ is open at zero.*

The test fails here, since no points of the form $(0, \varepsilon, *)$ belong to the image of the map

$$\begin{aligned}\mathbb{R}^5 \to \mathbb{R}^3 \colon (x_1, x_2, \theta, u_1, u_2)' &\mapsto f(x, u)\\ &= (u_1 \cos\theta, u_1 \sin\theta, u_2)'\end{aligned}$$

for $\theta \in (-\pi/2, \pi/2)$, unless $\varepsilon = 0$.

More generally, it is impossible to continuously stabilize any system without drift

$$\dot{x} = u_1 g_1(x) + \ldots + u_m g_m(x) = G(x)u$$

if $m < n$ and $\operatorname{rank}[g_1(0), \ldots, g_m(0)] = m$ (this includes all totally nonholonomic mechanical systems). Indeed, under these conditions, the map $(x, u) \mapsto G(x)u$ cannot contain a neighborhood of zero in its image, when restricted to a small enough neighborhood of zero. Indeed, let us first rearrange the rows of G:

$$G(x) \rightsquigarrow \begin{pmatrix} G_1(x) \\ G_2(x) \end{pmatrix}$$

so that $G_1(x)$ is of size $m \times m$ and is nonsingular for all states x that belong to some neighborhood N of the origin. Then,

$$\begin{pmatrix} 0 \\ a \end{pmatrix} \in \operatorname{Im}\left[N \times \mathbb{R}^m \to \mathbb{R}^n \colon (x, u) \mapsto G(x)u\right] \Rightarrow a = 0$$

(since $G_1(x)u = 0 \Rightarrow u = 0 \Rightarrow G_2(x)u = 0$ too).

If the condition $\operatorname{rank}[g_1(0), \ldots, g_m(0)] = m$ is violated, we cannot conclude a negative result. For instance, the system $\dot{x}_1 = x_1 u$, $\dot{x}_2 = x_2 u$ has $m = 1 < 2 = n$ but it can be stabilized by means of the feedback law $u = -(x_1^2 + x_2^2)$.

Observe that for linear systems, Brockett's condition says that

$$\operatorname{rank}[A, B] = n$$

which is the Hautus controllability condition (see e. g. [25], Lemma 3.3.7) at the zero mode.

Idea of the Proof One may prove Brockett's condition in several ways. A proof based on degree theory is probably easiest, and proceeds as follows (for details see for instance [25], Sect 5.9). The basic fact, due to Krasnosel'ski, is that if the system $\dot{x} = F(x) = f(x, k(x))$ has the origin as an asymptotically stable point and F is regular (since k is), then the degree (index) of F with respect to zero is $(-1)^n$, where n is the system dimension. In particular, the degree is also nonzero with respect to points p in a neighborhood of 0, which means that the equation $F(x) = p$ can be solved for small p, and hence $f(x, u) = p$ can be solved as well. The proof that the degree is $(-1)^n$ follows by exhibiting a homotopy, namely

$$F_t(x^0) = \frac{1}{t}\left[x\left(\frac{t}{1-t}, x^0\right) - x^0\right],$$

between $F_0 = F$ and $F_1(x) = -x$, and noting that the degree of the latter is obviously $(-1)^n$. An alternative proof uses Lyapunov functions. Asymptotic stability implies the existence of a smooth Lyapunov function V for $\dot{x} = F(x) = f(x, k(x))$, so, on the boundary ∂B of a sublevel set $B = \{x | V(x) \leq c\}$ we have that F points towards the interior of B. Thus, for p small, $F(x) - p$ still points to the interior, which means that B is invariant with respect to the perturbed vector field $\dot{x} = F(x) - p$. Provided that a fixed-point theorem applies to continuous maps $B \to B$, this implies that $F(x) - p$ must vanish somewhere in B, that is, the equation $F(x) = p$ can be solved. (Because, for each small $h > 0$, the time-h flow ϕ of $F - p$ has a fixed point $x_h \in B$, i. e. $\phi(h, x_h) = x_h$, so picking a convergent subsequence $x_h \to \bar{x}$ gives that $0 = (\phi(h, x_h) - x_h)/(h) \to F(\bar{x}) - p$.) A fixed point theorem can indeed be applied, because B is a retract of $\mathbb{R}^n$ (use the flow itself); note that this argument gives a weaker conclusion than the degree condition.

Control-Lyapunov Functions

The method of control-Lyapunov functions ("clf's") provides a powerful tool for studying stabilization problems, both as a basis of theoretical developments and as a method for actual feedback design.

Before discussing clf's, let us quickly review the classical concept of Lyapunov functions, through a simple example. Consider first a damped spring-mass system $\ddot{y} + \dot{y} + y = 0$, or, in state-space form with $x_1 = y$ and $x_2 = \dot{y}$, $\dot{x}_1 = x_2$, $\dot{x}_2 = -x_1 - x_2$. One way to verify global asymptotic stability of the equilibrium $x = 0$ is to pick the (Lyapunov) function $V(x_1, x_2) := \frac{3}{2}x_1^2 + x_1x_2 + x_2^2$, and observe that $\nabla V(x).f(x) = -|x|^2 < 0$ if $x \neq 0$, which means that

$$\frac{\mathrm{d}V(x(t))}{\mathrm{d}t} = -\left|x(t)^2\right| < 0$$

along all nonzero solutions, and thus the energy-like function V decreases along all trajectories, which, since V is a nondegenerate quadratic form, implies that $x(t)$ decreases, and in fact $x(t) \to 0$. Of course, in this case one could compute solutions explicitly, or simply note that the characteristic equation has all roots with negative real part, but Lyapunov functions are a general technique. (In fact, the classical converse theorems of Massera and Kurzweil [17,19] show that, whenever a system is GAS, there always exists a smooth Lyapunov function V.)

Now let us modify this example to deal with a control system, and consider a forced (but undamped) harmonic oscillator $\ddot{x} + x = u$, i. e. $\dot{x}_1 = x_2$, $\dot{x}_2 = -x_1 + u$. The damping feedback $u = -x_2$ stabilizes the system, but let us pretend that we do not know that. If we take the same V as before, now the derivatives along trajectories are, using "$\dot{V}(x, u)$" to denote $\nabla V(x).f(x, u)$ and omitting arguments t in $x(t)$ and $u(t)$:

$$\dot{V}(x, u) = -x_1^2 + x_1x_2 + x_2^2 - (x_1 + 2x_2)u\,.$$

This expression is affine in u. Thus, if x is a state such that $x_1 + 2x_2 \neq 0$, then we may pick a control value u (which depends on this current state x) such that $\dot{V} < 0$. On the other hand, if $x_1 + 2x_2 = 0$, then the expression reduces to $\dot{V} = -5x_2^2$ (for any u), which is negative unless x_2 (and hence also $x_1 = -2x_2$) vanishes.

In conclusion, for each $x \neq 0$ there is some u so that $\dot{V}(x, u) < 0$. This is, except for some technicalities to be discussed, the characterizing property of control-Lyapunov functions. For any given compact subset B in $\mathbb{R}^n$, we now pick some compact subset $\mathcal{U}_0 \subset \mathcal{U}$ so that

$$\forall x \in B, x \neq 0\,, \quad \exists u \in \mathcal{U}_0 \quad \text{such that} \quad \dot{V}(x, u) < 0\,. \tag{11}$$

In principle, then, we could then stabilize the system, for states in B, by using the steepest descent feedback law:

$$k(x) := \underset{u \in \mathcal{U}_0}{\operatorname{argmin}}\ \nabla V(x) \cdot f(x, u) \tag{12}$$

("argmin" means "pick any u at which the min is attained"; we restricted $\mathcal{U}$ to be assured that $\dot{V}(x, u)$ attains a minimum). Note that the stabilization problem becomes, in these terms, a set of static nonlinear programming problems: minimize a function of u, for each x. Global stabilization is also possible, by appropriately picking $\mathcal{U}_0$ as a function of the norm of x; later we discuss a precise formulation.

Control–Lyapunov functions, if understood non-technically as the basic paradigm "look for a function $V(x)$ with the properties that $V(x) \approx 0$ if and only if $x \approx 0$, and so that for each $x \neq 0$ it is possible to decrease $V(x)$ by some control action," constitute a very general approach to control (sometimes expressed in a dual fashion, as maximization of some measure of success). They appear in such disparate areas as A.I. game-playing programs (position evaluations), energy arguments for dissipative systems, program termination (Floyd/Dijkstra "variant"), and learning control ("critics" implemented by neural-networks). More relevantly to this paper, the idea underlies much of modern feedback control design, as illustrated for instance by the books [7,11,15,16,25].

Differentiable clf's: Precise Definition A *differentiable control-Lyapunov function* (clf) is a differentiable function $V\colon \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ which is proper, positive definite, and *infinitesimally decreasing*, meaning that there exists a positive definite continuous function $W\colon \mathbb{R}^n \to \mathbb{R}_{\geq 0}$, and there is some $\sigma \in \mathcal{N}$, so that

$$\sup_{x\in\mathbb{R}^n} \min_{|u|\leq\sigma(|x|)} \nabla V(x)\cdot f(x,u) + W(x) \leq 0\,. \tag{13}$$

This is basically the same as condition (11), with $\mathcal{U}_0 =$ the ball of radius $\sigma(|x|)$ picked as a function of x. The main difference is that, instead of saying "$\nabla V(x)\cdot f(x,u) < 0$ for $x \neq 0$" we write $\nabla V(x)\cdot f(x,u) \leq -W(x)$, where W is negative when $x \neq 0$. The two definitions are equivalent, but the "Hamiltonian" version used here is the correct one for the generalizations to be given, to nonsmooth V.

The basic result is due to Artstein [2]:

Theorem 4 *A control-affine system $\dot{x} = g_0(x) + \sum u_i g_i(x)$ admits a differentiable clf if and only if it admits a regular stabilizing feedback.*

The proof of sufficiency is easy: if there is such a k, then the converse Lyapunov theorem, applied to the closed-loop system $F(x) = f(x,k(x))$, provides a smooth V such that

$$L_F V(x) = \nabla V(x) F(x) < 0 \quad \forall x \neq 0\,.$$

This gives that for all nonzero x there is some u (bounded on bounded sets, because k is locally bounded by definition of feedback) so that $\dot{V}(x,u) < 0$; and one can put this in the form (13).

The necessity is more interesting. The original proof in [2] proceeds by a nonconstructive argument involving partitions of unity, but it is also possible [24,25] to exhibit explicitly a feedback, written as a function:

$$k\left(\nabla V(x)\cdot g_0(x), \ldots, \nabla V(x)\cdot g_m(x)\right)$$

of the directional derivatives of V along the vector fields defining the system (*universal formulas* for stabilization). Taking for simplicity $m = 1$, one such formula is:

$$k(x) := -\frac{a(x) + \sqrt{a(x)^2 + b(x)^4}}{b(x)} \quad (0 \text{ if } b = 0)$$

where $a(x) := \nabla V(x)\cdot g_0(x)$ and $b(x) := \nabla V(x)\cdot g_1(x)$. The expression for k is analytic in a,b when $x \neq 0$, because the clf property means that $a(x) < 0$ whenever $b(x) = 0$ [24,25].

Thus, the question of existence of regular feedback, for control-affine systems, reduces to the search for differentiable clf's, and this gives rise to a vast literature dealing with the construction of such V's, see [7,15,16,25] and references therein. Many other theoretical issues are also answered by Artstein's theorem. For example, via Kurzweil's converse theorem one has that the existence of k merely continuous on $\mathbb{R}^n \setminus \{0\}$ suffices for the existence of smooth (infinitely differentiable) V, and from here one may in turn find a k which is smooth on $\mathbb{R}^n \setminus \{0\}$. In addition, one may easily characterize the existence of k continuous at zero as well as regular: this is equivalent to the *small control property*: for each $\varepsilon > 0$ there is some $\delta > 0$ so that $0 < |x| < \delta$ implies that $\min_{|u|\leq\varepsilon} \nabla V(x)\cdot f(x,u) < 0$ (if this property holds, the universal formula automatically provides such a k). We should note that Artstein provided a result valid for general, not necessarily control-affine systems $\dot{x} = f(x,u)$; however, the obtained "feedback" has values in sets of relaxed controls, and is not a feedback law in the classical sense. Later, we discuss a different generalization.

Differentiable clf's will in general not exist, because of obstructions to regular feedback stabilization. This leads us naturally into the twin subjects of discontinuous feedbacks and non-differentiable clf's.

Discontinuous Feedback

The previous results and examples show that, in order to develop a satisfactory general theory of stabilization, one in which one proves the implication "asymptotic controllability implies feedback stabilizability," we must allow discontinuous feedback laws $u = k(x)$. But then, a major technical difficulty arises: solutions of the initial-value problem $\dot{x} = f(x,k(x))$, $x(0) = x^0$, interpreted in the classical sense of differentiable functions or even as (absolutely) continuous solutions of the integral equation $x(t) = x^0 + \int_0^t f(x(s), k(x(s)))\mathrm{d}s$, do not exist in general. The only general theorems apply to systems $\dot{x} = F(x)$ with continuous F. For example, there is no solution to $\dot{x} = -\operatorname{sign} x$, $x(0) = 0$, where $\operatorname{sign} x = -1$ for $x < 0$ and $\operatorname{sign} x = 1$ for $x \geq 0$. So one cannot even pose the stabilization problem in a mathematically consistent sense.

There is, of course, an extensive literature addressing the question of discontinuous feedback laws for control systems and, more generally, differential equations with discontinuous right-hand sides. One of the best-known candidates for the concept of solution of (3) is that of a *Filippov solution* [6,9], which is defined as the solution of a certain differential inclusion with a multivalued right-hand side which is built from $f(x, k(x))$. Unfortunately, there is no hope of obtaining the implication "asymptotic controllability implies feedback stabilizability" if one interprets solutions of (3) as Filippov solutions. This is a consequence of results in [5,22], which established that the existence of a discontinuous stabilizing feedback in the Filippov sense implies the Brockett necessary conditions, and, moreover, for systems affine in controls it also implies the existence of regular feedback (which we know is in general impossible).

A different concept of solution originates with the theory of discontinuous positional control developed by Krasovskii and Subbotin in the context of differential games in [14], and it is the basis of the new approach to discontinuous stabilization proposed in [4], to which we now turn.

Limits of High-Frequency Sampling

By a *sampling schedule* or *partition* $\pi = \{t_i\}_{i \geq 0}$ of $0, +\infty$ we mean an infinite sequence

$$0 = t_0 < t_1 < t_2 < \ldots$$

with $\lim_{i\to\infty} t_i = \infty$. We call

$$\mathbf{d}(\pi) := \sup_{i \geq 0}(t_{i+1} - t_i)$$

the *diameter* of π. Suppose that k is a given feedback law for system (1). For each π, the *π-trajectory starting from x^0* of system (3) is defined recursively on the intervals $[t_i, t_{i+1})$, $i = 0, 1, \ldots$, as follows. On each interval $t_i, t_{i+1})$, the initial state is measured, the control value $u_i = k(x(t_i))$ is computed, and the constant control $u \equiv u_i$ is applied until time t_{i+1}; the process is then iterated. That is, we start with $x(t_0) = x^0$ and solve recursively

$$\dot{x}(t) = f(x(t), k(x(t_i)))\,, t \in t_i, t_{i+1})\,, \quad i = 0, 1, 2, \ldots$$

using as initial value $x(t_i)$ the endpoint of the solution on the preceding interval. The ensuing π-trajectory, which we denote as $x_\pi(\cdot, x^0)$, is defined on some maximal nontrivial interval; it may fail to exist on the entire interval $[0, +\infty)$ due to a blow-up on one of the subintervals $t_i, t_{i+1})$. We say that it is *well defined* if $x_\pi(t, x^0)$ is defined on all of $[0, +\infty)$.

Definition The feedback $k \colon \mathbb{R}^n \to \mathcal{U}$ *stabilizes* the system (1) if there exists a function $\beta \in \mathcal{KL}$ so that the following property holds: For each

$$0 < \varepsilon < K$$

there exists a $\delta = \delta(\varepsilon, K) > 0$ such that, for every sampling schedule π with $\mathbf{d}(\pi) < \delta$, and for each initial state x^0 with $|x^0| \leq K$, the corresponding π-trajectory of (3) is well-defined and satisfies

$$\left|x_\pi(t, x^0)\right| \leq \max\left\{\beta\left(K, t\right), \varepsilon\right\} \quad \forall t \geq 0\,. \tag{14}$$

In particular, we have

$$\left|x_\pi(t, x^0)\right| \leq \max\left\{\beta\left(\left|x^0\right|, t\right), \varepsilon\right\} \quad \forall t \geq 0 \tag{15}$$

whenever $0 < \varepsilon < |x^0|$ and $\mathbf{d}(\pi) < \delta(\varepsilon, |x^0|)$ (just take $K := |x^0|$).

Observe that the role of δ is to specify a lower bound on intersampling times. Roughly, one is requiring that

$$t_{i+1} \leq t_i + \theta\left(|x(t_i)|\right)$$

for each i, where θ is an appropriate positive function.

Our definition of stabilization is physically meaningful, and is very natural in the context of sampled-data (computer control) systems. It says in essence that a feedback k stabilizes the system if it drives all states asymptotically to the origin and with small overshoot when using *any fast enough sampling schedule*. A high enough sampling frequency is generally required when close to the origin, in order to guarantee small displacements, and also at infinity, so as to preclude large excursions or even blow-ups in finite time. This is the reason for making δ depend on ε and K.

This concept of stabilization can be reinterpreted in various ways. One is as follows. Pick any initial state x^0, and consider any sequence of sampling schedules π_ℓ whose diameters $\mathbf{d}(\pi_\ell)$ converge to zero as $\ell \to \infty$ (for instance, constant sampling rates with $t_i = i/\ell$, $i = 0, 1, 2, \ldots$). Note that the functions $x_\ell := x_{\pi_\ell}(\cdot, x^0)$ remain in a bounded set, namely the ball of radius $\beta(|x^0|, 0)$ (at least for ℓ large enough, for instance, any ℓ so that $\mathbf{d}(\pi_\ell)\delta(|x^0|/2, |x^0|)$). Because $f(x, k(x))$ is bounded on this ball, these functions are equicontinuous, and (Arzela–Ascoli's Theorem) we may take a subsequence, which we denote again as $\{x_\ell\}$, so that $x_\ell \to x$ as $\ell \to \infty$ (uniformly on compact time intervals) for some absolutely continuous (even Lipschitz) function $x\colon [0,\infty) \to \mathbb{R}^n$. *We may think of any limit function $x(\cdot)$ that arises in this fashion as a generalized solution of the closed-loop Eq.* (3). That is, generalized solutions are the limits of trajectories arising from arbitrarily high-frequency sampling when using the feedback law $u = k(x)$. Generalized solutions, for a given initial state x^0, may not be unique – just as may happen with continuous but non-Lipschitz feedback – but there is always existence, and, moreover, for any generalized solution, $|x(t)| \leq \beta(|x^0|\,, t)$ for all $t \geq 0$. This is precisely the defining estimate for the GAS property. Moreover, if k happens to be regular, then the unique solution of $\dot{x} = f(x, k(x))$ in the classical sense is also the unique generalized solution, so we have a reasonable extension of the concept of solution. (This type of interpretation is somewhat analogous, at least in spirit, to the way in which "relaxed" controls are interpreted in optimal trajectory calculations, namely through high-frequency switching of approximating regular controls.) The definition of stabilization was given in [4] in a slightly different form; see [26] for a discussion of the equivalence.

Stabilizing Feedbacks Exist

The main result is [4]:

Theorem 5 *The system (1) admits a stabilizing feedback if and only if it is asymptotically controllable.*

Necessity is clear. The sufficiency statement is proved by construction of k, and is based on the following ingredients:

- Existence of a nonsmooth control-Lyapunov function V.
- Regularization on shells of V.
- Pointwise minimization of a Hamiltonian for the regularized V.

In order to sketch this construction, we start by quickly reviewing a basic concept from nonsmooth analysis.

Proximal Subgradients Let V be any continuous function $\mathbb{R}^n \to \mathbb{R}$ (or even, just lower semicontinuous and with extended real values). A *proximal subgradient* of V at the point $x \in \mathbb{R}^n$ is any vector $\zeta \in \mathbb{R}^n$ such that, for some $\sigma > 0$ and some neighborhood $\mathcal{O}$ of x,

$$V(y) \geq V(x) + \zeta \cdot (y - x) - \sigma^2 \left|y - x^2\right| \quad \forall y \in \mathcal{O}\,.$$

In other words, proximal subgradients are the possible gradients of supporting quadratics at the point x. The set of all proximal subgradients at x is denoted $\partial_p V(x)$.

Nonsmooth Control–Lyapunov Functions A continuous (but not necessarily differentiable) $V\colon \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ is a *control-Lyapunov function* (clf) if it is proper, positive definite, and infinitesimally decreasing in the following generalized sense: there exist a positive definite continuous $W\colon \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ and a $\sigma \in \mathcal{N}$ so that

$$\sup_{x \in \mathbb{R}^n} \max_{\zeta \in \partial_p V(x)} \min_{|u| \leq \sigma(|x|)} \zeta \cdot f(x, u) + W(x) \leq 0\,. \quad (16)$$

This is the obvious generalization of the differentiable case in (13); we are still asking that one should be able to make $\nabla V(x) \cdot f(x, u) < 0$ by an appropriate choice of $u = u_x$, for each $x \neq 0$, except that now we replace $\nabla V(x)$ by the proximal subgradient set $\partial_{\mathrm{p}} V(x)$. An equivalent property is to ask that V be a viscosity supersolution of the corresponding Hamilton–Jacobi–Bellman equation.

For nonsmooth clf's, the main basic result is [23,26]:

Theorem 6 *The system (1) is asymptotically controllable if and only if it admits a continuous clf.*

The proof is based on first constructing an appropriate W, and then letting V be the optimal cost (Bellman function) for the problem min $\int_0^\infty W(x(s))\mathrm{d}s$. However, some care has to be taken to insure that V is continuous, and the cost has to be adjusted in order to deal with possibly unbounded minimizers. An important and very useful refinement of this result is the fact that a locally Lipschitz clf can also be shown to exist [21].

Regularization Once V is known to exist, the next step in the construction of a stabilizing feedback is to obtain Lipschitz approximations of V. For this purpose, one considers the Iosida–Moreau inf-convolution of V with a quadratic function:

$$V_\alpha(x) := \inf_{y \in \mathbb{R}^n} \left[V(y) + \frac{1}{2\alpha^2} |y - x|^2 \right]$$

where the number $\alpha > 0$ is picked constant on appropriate regions. One has that $V_\alpha(x) \nearrow V(x)$, uniformly on compacts. Since V_α is locally Lipschitz, Rademacher's Theorem insures that V_α is differentiable almost everywhere. The feedback k is then made equal to a pointwise minimizer k_α of the Hamiltonian, at the points of differentiability (compare with (12) for the case of differentiable V):

$$k_\alpha(x) := \operatorname*{argmin}_{u \in \mathcal{U}_0} \nabla V_\alpha(x) \cdot f(x, u)\,,$$

where α and the compact $\mathcal{U}_0 = \mathcal{U}_0(\alpha)$ are chosen constant on certain compacts and this choice is made in between level curves. The critical fact is that V_α is itself a clf for the original system, at least when restricted to the region where it is needed. More precisely, on each shell of the form

$$C = \{x \in \mathbb{R}^n | r \leq |x| \leq R\} ,$$

there are positive numbers m and α_0 and a compact subset $\mathcal{U}_0$ such that, for each $0 < \alpha \leq \alpha_0$, each $x \in C$, and every $\zeta \in \partial_{\mathrm{p}} V_\alpha(x)$,

$$\min_{u \in \mathcal{U}_0} \zeta \cdot f(x, u) + m \leq 0 .$$

See [26]. (Actually, this description is oversimplified, and the proof is a bit more delicate. One must define, on appropriate compact sets

$$k(x) := \operatorname*{argmin}_{u \in \mathcal{U}_0} \zeta_\alpha(x) \cdot f(x, u) ,$$

where $\zeta_\alpha(x)$ is carefully chosen. At points x of nondifferentiability, $\zeta_\alpha(x)$ is not a proximal subgradient of V_α, since $\partial_{\mathrm{p}} V_\alpha(x)$ may well be empty. One uses, instead, the fact that $\zeta_\alpha(x)$ happens to be in $\partial_{\mathrm{p}} V(x')$ for some $x' \approx x$. See [4] for details.)

Sensitivity to Small Measurement Errors

We have seen that every asymptotically controllable system admits a feedback stabilizer k, generally discontinuous, which renders the closed-loop system $\dot{x} = f(x, k(x))$ GAS. On the other hand, one of the main motivations for the use of feedback is in order to deal with uncertainty, and one possible source of uncertainty are measurement errors in state estimation. The use of discontinuous feedback means that undesirable behavior – chattering – may arise. In fact, one of the main reasons for the focus on continuous feedback is precisely in order to avoid such behaviors. Thus, we turn now to an analysis of the effect of measurement errors. Suppose first that k is a continuous function of x. Then, if the error e is small, using the control $u' = k(x + e)$ instead of $u = k(x)$ results in behavior which remains close to the intended one, since $k(x + e) \approx k(x)$; moreover, if $e \ll x$ then stability is preserved. This property of robustness to small errors when k is continuous can be rigorously established by means of a Lyapunov proof, based on the observation that, if V is a Lyapunov function for the closed-loop system, then continuity of $f(x, k(x + e))$ on e means that

$$\nabla V(x) \cdot f(x, k(x + e)) \approx \nabla V(x) \cdot f(x, k(x)) < 0 .$$

Unfortunately, when k is not continuous, this argument breaks down. However, it can be modified so as to avoid invoking continuity of k. Assuming that V is continuously differentiable, one can argue that

$$\nabla V(x) \cdot f(x, k(x+e)) \approx \nabla V(x+e) \cdot f(x, k(x+e)) < 0$$

(using the Lyapunov property at the point $x + e$ instead of at x). This observation leads to a theorem, formulated below, which says that a discontinuous feedback stabilizer, robust with respect to small observation errors, can be found provided that there is a C^1 clf. In general, as there are no C^1, but only continuous, clf's, one may not be able to find any feedback law that is robust in this sense.

There are many well-known techniques for avoiding chattering, and a very common one is the introduction of deadzones where no action is taken. The feedback constructed in [4], with no modifications needed, can always be used in a manner robust with respect to small observation errors, using such an approach. Roughly speaking, the general idea is as follows. Suppose that the true current state, let us say at time $t = t_i$, is x, but that the controller uses $u = k(\tilde{x})$, where $\tilde{x} = x + e$, and e is small. Call x' the state that results at the next sampling time, $t = t_{i+1}$. By continuity of solutions on initial conditions, $|x' - \tilde{x}'|$ is also small, where $\tilde{x}'$ is the state that would have resulted from applying the control u if the true state had been $\tilde{x}$. By continuity, it follows that $V_\alpha(x) \approx V_\alpha(\tilde{x})$ and also $V_\alpha(x') \approx V_\alpha(\tilde{x}')$. On the other hand, the construction in [4] provides that $V_\alpha(\tilde{x}') < V_\alpha(\tilde{x}) - \mathrm{d}(t_{i+1} - t_i)$, where d is some positive constant (this is valid while we are far from the origin). Hence, if e is sufficiently small compared to the intersample time $t_{i+1} - t_i$, it will necessarily be the case that $V_\alpha(x')$ must also be smaller than $V_\alpha(x)$. This discussion may be formalized in several ways; see [26] for a precise statement.

If we insist upon fast sampling, a necessary condition arises, as was proved in the recent paper [18] (which, in turn, represented an extension of the work by Hermes [10] for classical solutions under observation error). We next discuss the main result from that paper. We consider systems

$$\dot{x}(t) = f(x(t), k(x(t) + \mathrm{e}(t)) + \mathrm{d}(t)) \qquad (17)$$

in which there are observation errors as well as, now, possible actuator errors $d(\cdot)$. Actuator errors $d(\cdot)\colon [0, \infty) \to \mathcal{U}$ are Lebesgue measurable and locally essentially bounded, and observation errors $\mathrm{e}(\cdot)\colon [0, \infty) \to \mathbb{R}^n$ are locally bounded. We define solutions of (17), for each sampling schedule π, in the usual manner, i. e., solving recursively on the intervals $t_i, t_{i+1})$, $i = 0, 1, \ldots$, the differential equation

$$\dot{x}(t) = f\big(x(t), k(x(t_i) + \mathrm{e}(t_i)) + \mathrm{d}(t)\big) \qquad (18)$$

with $x(0) = x^0$. We write $x(t) = x_\pi(t, x^0, \mathrm{d}, \mathrm{e})$ for the solution, and say it is *well-defined* if it is defined for all $t \geq 0$.

Definition The feedback $k\colon \mathbb{R}^n \to \mathcal{U}$ *stabilizes* the system (17) if there exists a function $\beta \in \mathcal{KL}$ so that the following property holds: For each

$$0 < \varepsilon < K$$

there exist $\delta = \delta(\varepsilon, K) > 0$ and $\eta = \eta(\varepsilon, K)$ such that, for every sampling schedule π with $\mathbf{d}(\pi) < \delta$, each initial state x^0 with $|x^0| \leq K$, and each e, d such that $|\mathrm{e}(t)| \leq \eta$ for all $t \geq 0$ and $|\mathrm{d}(t)| \leq \eta$ for almost all $t \geq 0$, the corresponding π-trajectory of (17) is well-defined and satisfies

$$\left|x_\pi(t, x^0, \mathrm{d}, \mathrm{e})\right| \leq \max\left\{\beta\left(K, t\right), \varepsilon\right\} \quad \forall t \geq 0\,. \qquad (19)$$

In particular, taking $K := |x^0|$, one has that

$$\left|x_\pi(t, x^0, d, e)\right| \leq \max\left\{\beta\left(\left|x^0\right|, t\right), \varepsilon\right\} \quad \forall t \geq 0$$

whenever $0 < \varepsilon < |x^0|$, $\mathbf{d}(\pi) < \delta(\varepsilon, |x^0|)$, and for all t, $|\mathrm{e}(t)| \leq \eta(\varepsilon, |x^0|)$, and $|\mathrm{d}(t)| \leq \eta(\varepsilon, |x^0|)$.

The main result in [18] is as follows.

Theorem 7 *There is a feedback which stabilizes the system (17) if and only if there is a C^1 clf for the unperturbed system (1).*

It is interesting to note that, as a corollary of Artstein's Theorem, for control-affine systems $\dot{x} = g_0(x) + \sum u_i\, g_i(x)$ we may conclude that if there is a discontinuous feedback stabilizer that is robust with respect to small noise, then there is also a regular one, and even one that is smooth on $\mathbb{R}^n \setminus \{0\}$. For non control-affine systems, however, there may exist a discontinuous feedback stabilizer that is robust with respect to small noise, yet there is no regular feedback.

Briefly, the sufficiency part of Theorem 7 proceeds by taking a pointwise minimization of the Hamiltonian, for a given C^1 clf, i. e. $k(x)$ is defined as any u with $|u| \leq \sigma(|x|)$ which minimizes $\nabla V(x) \cdot f(x, u)$. The necessity part is based on the following technical fact: if the perturbed system can be stabilized, then the differential inclusion

$$\dot{x} \in F(x) := \bigcap_{\varepsilon > 0} \overline{\mathrm{co}}\, f(x, k(x + \varepsilon B))$$

(where B denotes the unit ball in $\mathbb{R}^n$) is strongly asymptotically stable. One may then apply converse Lyapunov theorems for upper semicontinuous compact convex differential inclusions to deduce the existence of V.

We now summarize exactly which implications hold, writing "robust" to mean stabilization of the system subject to observation and actuator noise:

$$\begin{tikzcd}
C^1 V \arrow[r, Leftrightarrow] \arrow[d, Rightarrow] & \exists \text{ robust } k \arrow[d, Rightarrow] & \\
C^0 V \arrow[r, Leftrightarrow] & \exists k \arrow[r, Leftrightarrow] & \mathrm{AC}\,.
\end{tikzcd}$$

Future Directions

There are several alternative approaches to feedback stabilization, notably the very appealing approach to discontinuous stabilization throgh *patchy feedbacks* [1], as well as other related "hybrid" approaches [20]. It is also extremely important to understand the effect of "large" disturbances on the behavior of feedback systems. This study leads one to the very active area of input to state stability (ISS) and related notions (output to state stability as a model of detectability, input to output stability for the study of regulation problems, and so forth), see [27].

Bibliography

Primary Literature

1. Ancona F, Bressan A (2003) Flow stability of patchy vector fields and robust feedback stabilization. SIAM J Control Optim 41:1455–1476
2. Artstein Z (1983) Stabilization with relaxed controls. Nonlinear Anal Theory Methods Appl 7:1163–1173
3. Brockett R (1983) Asymptotic stability and feedback stabilization. In: Differential geometric control theory. Birkhauser, Boston, pp 181–191
4. Clarke FH, Ledyaev Y, Sontag E, Subbotin A (1997) Asymptotic controllability implies feedback stabilization. IEEE Trans Autom Control 42(10):1394–1407
5. Coron J-M, Rosier L (1994) A relation between continuous time-varying and discontinuous feedback stabilization. J Math Syst Estim Control 4:67–84
6. Filippov A (1985) Differential equations with discontinuous right-hand side. Nauka, Moscow. English edition: Kluwer, Dordrecht, (1988)
7. Freeman R, Kokotović P (1996) Robust nonlinear control design. Birkhäuser, Boston
8. Grüne L, Sontag E, Wirth F (1999) Asymptotic stability equals exponential stability, and ISS equals finite energy gain—if you twist your eyes. Syst Control Lett 38(2):127–134
9. Hájek O (1979) Discontinuous differential equations, i & ii. J Diff Equ 32(2):149–170, 171–185
10. Hermes H (1967) Discontinuous vector fields and feedback control. In: Differential equations and dynamical systems. Proc Internat Sympos, Mayaguez 1965. Academic Press, New York, pp 155–165

11. Isidori A (1995) Nonlinear control systems: An introduction. Springer, Berlin
12. Kawski M (1989) Stabilization of nonlinear systems in the plane. Syst Control Lett 12:169–175
13. Kawski M (1995) Geometric homogeneity and applications to stabilization. In: Krener A, Mayne D (eds) Nonlinear control system design symposium (NOLCOS), Lake Tahoe, June 1995. Elsevier
14. Krasovskii N, Subbotin A (1988) Game-theoretical control problems. Springer, New York
15. Krstić M, Deng H (1998) Stabilization of uncertain nonlinear systems. Springer, London
16. Krstić M, Kanellakopoulos I, Kokotović PV (1995) Nonlinear and adaptive control design. Wiley, New York
17. Kurzweil J (1956) On the inversion of Ljapunov's second theorem on stability of motion. Am Math Soc Transl 2:19–77
18. Ledyaev Y, Sontag E (1999) A Lyapunov characterization of robust stabilization. Nonlinear Anal 37(7):813–840
19. Massera JL (1956) Contributions to stability theory. Ann Math 64:182–206
20. Prieur C (2006) Robust stabilization of nonlinear control systems by means of hybrid feedbacks. Rend Sem Mat Univ Pol Torino 64:25–38
21. Rifford L (2002) Semiconcave control-Lyapunov functions and stabilizing feedbacks. SIAM J Control Optim 41:659–681
22. Ryan E (1994) On Brockett's condition for smooth stabilizability and its necessity in a context of nonsmooth feedback. SIAM J Control Optim 32:1597–1604
23. Sontag E (1983) A Lyapunov-like characterization of asymptotic controllability. SIAM J Control Optim 21(3):462–471
24. Sontag E (1989) A "universal" construction of Artstein's theorem on nonlinear stabilization. Syst Control Lett 13(2):117–123
25. Sontag E (1998) Mathematical control theory. In: Deterministic finite-dimensional systems. Texts in Applied Mathematics, vol 6. Springer, New York
26. Sontag E (1999) Stability and stabilization: Discontinuities and the effect of disturbances. In : Nonlinear analysis, differential equations and control (Montreal, 1998). NATO Sci Ser C Math Phys Sci, vol 528. Kluwer, Dordrecht, pp 551–598
27. Sontag E (2006) Input to state stability: Basic concepts and results. In: Nistri P, Stefani G (eds) Nonlinear and optimal control theory. Springer, Berlin, pp 163–220
28. Sontag E, Sussmann H (1980) Remarks on continuous feedback. In: Proc IEEE Conf Decision and Control, Albuquerque, Dec 1980. pp 916–921

Books and Reviews

Clarke FH, Ledyaev YS, Stern RJ, Wolenski P (1998) Nonsmooth analysis and control theory. Springer, New York

Freeman R, Kokotović PV (1996) Robust nonlinear control design. Birkhäuser, Boston

Isidori A (1995) Nonlinear control systems, 3rd edn. Springer, London

Isidori A (1999) Nonlinear control systems II. Springer, London

Khalil HK (1996) Nonlinear systems, 2nd edn. Prentice-Hall, Upper Saddle River

Krstić M, Deng H (1998) Stabilization of uncertain nonlinear systems. Springer, London

Sepulchre R, Jankovic M, Kokotović PV (1997) Constructive nonlinear control. Springer, New York

Stability Theory of Ordinary Differential Equations

Carmen Chicone
Department of Mathematics, University of Missouri-Columbia, Columbia, USA

Article Outline

Glossary

Ordinary differential equation An equation for an unknown vector of functions of a single variable that involves derivatives of the unknown functions. The order of a differential equation is the highest order of the derivatives that appear. The most important class of differential equations are first-order systems of ordinary differential equations that can be written in the form $\dot{u} = f(u, t)$, where f is a given smooth function $f\colon U \times J \to \mathbb{R}^n$, U is an open subset of $\mathbb{R}^n$, and J is an open subset of $\mathbb{R}$. The unknown functions are the components of u, and the vector of their first-order derivatives with respect to the independent variable t is denoted by $\dot{u}$. A solution of this differential equation is a function $u\colon K \to \mathbb{R}^n$, where K is an open subset of J such that $\frac{\mathrm{d}u}{\mathrm{d}t}(t) = f(u(t), t)$ for all $t \in K$.

Dynamical system A set and a law of evolution for its elements. The first-order differential equation $\dot{u} = f(u, t)$, where $f\colon U \times J \to \mathbb{R}^n$, is the law of evolution for the set $U \times J$; it defines a continuous (time) dynamical system: Given $(v, s) \in U \times J$, the solution $t \mapsto \psi(t, s, v)$ such that $\phi(s, s, v) = v$ determines the evolution of the state v: the state v at time s evolves to the state $\psi(t, s, v)$ at time t. Similarly, a continuous function $f\colon X \to X$ on a metric space X defines a discrete dynamical system. The state $x \in X$ evolves to $f^k(x)$ (which denotes the value of f composed with itself k times and evaluated at x) after k time-steps. The images of $t \mapsto \phi(t, s, v)$ and $k \mapsto f^k(x)$ are called the orbits of the corresponding states v and x.

Stability theory The mathematical analysis of the behavior of the distances between an orbit (or set of orbits) of a dynamical system and all other nearby orbits.

Definition of the Subject

An orbit or set of orbits of a dynamical system is stable if all solutions starting nearby remain nearby for all future times. This concept is of fundamental importance in applied mathematics: the stable solutions of mathematical models of physical processes correspond to motions that are observed in nature.

Introduction

Stability theory began with a basic question about the natural world: Is the solar system stable? Will the present configuration of the planets and the sun remain forever; or, might some planets collide, radically change their orbits, or escape from the solar system?

With the advent of Isaac Newton's second law of motion and the law of universal gravitation, the motions of the planets in the solar system were understood to correspond to the solutions of the Newtonian system of ordinary differential equations that modeled the positions and velocities of the planets and the sun according to their mutual gravitational attractions. Short-term approximate predictions (up to a few years in the future) verified this theory; but, due to the complexity of the differential equations of motion, the problem of long-term prediction was not solved; it still occupies a central place in mathematical research.

Does the Newtonian model predict the stability of the solar system?

During the 18th Century, Pierre Simon Laplace [52, 53] asserted a proof of the stability of the solar system. He considered the changes in the semi-major axes and eccentricities of the elliptical motions of the planets around the sun. Using (reasonable) approximations of the Newtonian equations of motion, he showed that for his approximate model these orbital elements do not change over long periods of time due to the disturbances caused by the gravitational attractions of the other bodies in the solar system. If true for the full Newtonian equations of motion, these assertions would imply the stability of the Newtonian solar system. In fact, no proof of the stability of the solar system is known (see [69]). Laplace's results merely provide evidence in favor of stability of the solar system; on the other hand, this work was a primary stimulus for the later development of a general theory of stability.

One of the first rigorous results in stability theory was stated by Joseph-Louis Lagrange [49] and proved by Lejeune Dirichlet [23,24]; it states that an isolated minimum of the potential energy of a conservative mechanical system is the position of a stable equilibrium point.

Joseph Liouville [56,57,58,59] discussed the problem of the stability of rotating fluid bodies. The further development of this theory was suggested to Aleksandr Mikhailovich Lyapunov as a thesis topic by his advisor Pafnuty Chebyshev. This led to the fundamental and foundational work of Lyapunov on stability theory [63].

Henri Poincaré's introduction of the qualitative theory of differential equations [71] influenced Lyapunov's treatment of stability theory and laid much of the foundation for the modern theory of nonlinear dynamical systems. In addition, Poincaré's work on celestial mechanics [72,73,74,75,76] discusses stability theory.

Mathematical Formulation of the Stability Concept and Basic Results

Consider the first-order ordinary differential equation

$$\dot{u} = f(u, t) , \tag{1}$$

where the dependent variable u is an n-dimensional real vector, $\dot{u}$ denotes the derivative of u with respect to the independent variable t, and f (a mapping, from the cross product of an open subset U of n-dimensional space $\mathbb{R}^n$ and an open subset J of the real line $\mathbb{R}$, into $\mathbb{R}^n$) is at least class C^1. The existence and uniqueness theorem for differential equations (see, for example, [19]) states that if $(v, s) \in U \times J$, then there is a unique solution $t \mapsto \psi(t, s, v)$ of the differential equation such that $\psi(s, s, v) = v$. Moreover, the function ψ is as smooth as the function f.

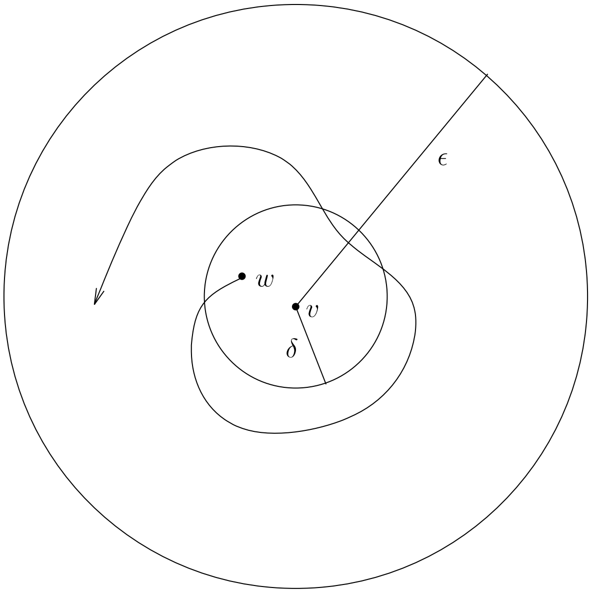
Definition 1 For $v \in \mathbb{R}^n$, the notation $|v|$ denotes the length of v.

A solution $t \mapsto \psi(t, s, v)$ of the differential Eq. (1) is called stable if it is defined for all time $t \geq s$ and for every positive number ϵ there is a positive number δ such that $|\phi(t, s, w) - \phi(t, s, v)| < \epsilon$ whenever $t \geq s$ and $|w - v| < \delta$ (see Fig. 1).

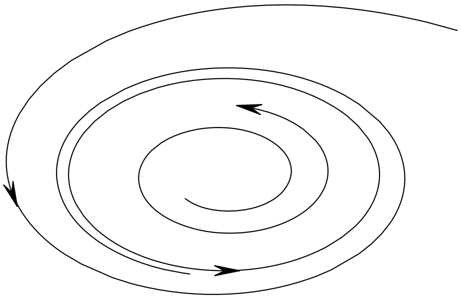
A stable solution $t \mapsto \psi(t, s, v)$ is called orbitally asymptotically stable if there is a choice of δ such that the distance between the point $\psi(t, s, w)$ and the set $\{\psi(t, s, v) \colon t \geq s\}$—called the forward orbit of the solution—converges to zero as $t \to \infty$.

A stable solution $t \mapsto \psi(t, s, v)$ is called asymptotically stable if there is a choice of δ such that the distance $|\phi(t, s, w) - \phi(t, s, v)|$ converges to zero as $t \to \infty$ whenever $|w - v| < \delta$ (see Fig. 2).

A solution is called unstable if it is not stable.



Stability Theory of Ordinary Differential Equations, Figure 1
The figure depicts a stable rest point *v*. The orbit starting at *w*, an arbitrary point whose distance from *v* is less than δ, remains within distance ϵ for all positive time



Stability Theory of Ordinary Differential Equations, Figure 2
The figure depicts an orbitally asymptotically stable elliptical periodic orbit together with two additional orbits that approach the periodic orbit, one from the outside and one from the inside of the region bounded by the periodic orbit

While the definitions just given are widely accepted, the words “asymptotic stability” are often used to mean orbital asymptotic stability. For most situations orbital asymptotic stability is the desired concept. It is clear from the definitions that the concept of asymptotic stability is much stronger than the concept of orbital asymptotic stability for stable solutions whose orbits consist of more than one point. The position of the evolving state along an asymptotically stable orbit is approached by an open set of evolving states. On the other hand, for a solution to be orbitally asymptotically stable only the distances between the elements of this open set of evolving states and the point set consisting of the forward orbit of the stable solution is required to approach zero. There is no requirement that the open set of evolving states all converge to the evolving state on the stable orbit in unison as time increases without bound. An important concept introduced by Poincaré, the return map, is useful in the study of orbital asymptotic stability; it will be discussed below.

To illustrate the concept of stability, let us consider the (dimensionless) harmonic oscillator

$$\ddot{x} + x = 0 \,. \tag{2}$$

This second-order differential equation is recast as the first-order system

$$\dot{x} = y \,, \quad \dot{y} = -x \tag{3}$$

by introducing the velocity $y := \dot{x}$ as a new variable. It is a classical mechanical system with kinetic energy $\frac{1}{2}\dot{x}^2$ and potential energy $\frac{1}{2}x^2$. According to the principle of Lagrange, the equilibrium solution $(x, y) = (0, 0)$ (which corresponds to an isolated minimum of the potential energy) is stable. This result is easily verified by inspection of the general solution of system (3) that is given by

$$\phi(t, \xi, \eta) = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix} \begin{pmatrix} \xi \\ \eta \end{pmatrix} , \tag{4}$$

where (ξ, η) is an arbitrary point in $\mathbb{R}^2$. All non-equilibrium solutions are periodic. In fact, the orbit of the solution starting at (ξ, η) is a circle with radius $\sqrt{\xi^2 + \eta^2}$. Given an open subset containing the origin, the disk bounded by one of these circles defines an open subset. Every orbit with initial value inside this disk remains inside the disk (and hence the given open set) for all $t \geq 0$. On the other hand, the origin is not orbitally asymptotically stable; indeed, the periodic solutions do not approach the origin in positive time.

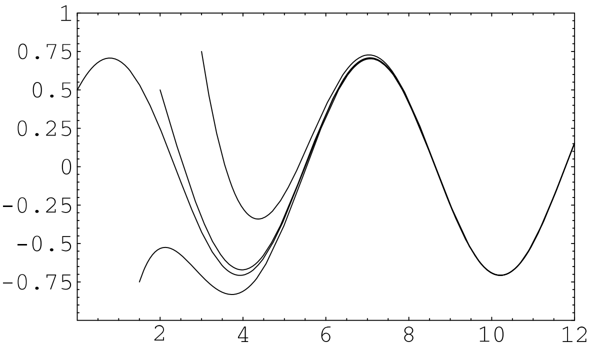
The origin is also an equilibrium of the damped harmonic oscillator

$$\ddot{x} + \epsilon\dot{x} + x = 0 \,, \tag{5}$$

where $\epsilon > 0$. Again, inspection of the general solution shows that the origin is asymptotically stable and therefore orbitally asymptotically stable. In effect, all terms in the formula for the general solution contain the factor $\mathrm{e}^{-\epsilon t/2}$, which converges to zero as $t \to \infty$.

For a nonautonomous example (that is, a differential Eq. (1) such that the partial derivative of f with respect to t is not zero), consider the first-order scalar differential equation

$$\dot{x} = -x + \sin t \tag{6}$$



Stability Theory of Ordinary Differential Equations, Figure 3
A graph of the asymptotically stable solution (7) of the scalar first-order differential Eq. (6) is depicted together with the graphs of three additional solutions given by $t \mapsto \phi(t,s,v)$ for (s,v) equal to $(1.5,-0.75)$, $(3.0,0.75)$, and $(2.0,0.5)$

whose solutions are given by

$$\phi(t,s,v) = \left(v - \frac{1}{2}(\sin s + \cos s)\right) e^{s-t} + \frac{1}{2}(\sin t + \cos t)\,.$$

The solution starting at the state 1/2 at time zero,

$$\phi\left(t, 0, \frac{1}{2}\right) = \frac{1}{2}(\sin t + \cos t)\,, \tag{7}$$

is asymptotically stable. This result is true due to the presence of the exponential factor in the general solution; it decreases to zero for each fixed s as $t \to \infty$ (cf. Fig. 3).

The origin is an unstable equilibrium for the scalar first-order differential equation $\dot{x} = x^2$, whose solutions are given by

$$\phi(t,s,v) = \frac{v}{1-(t-s)v}\,.$$

Indeed, for every $\delta > 0$ (no matter how small) there is a some $v \in \mathbb{R}$ such that $|v| < \delta$ and $v > 0$. The corresponding solution converges to infinity as $t \to (1+sv)/v$.

Because explicit solutions of differential equations are rare, the main subject of stability theory is the determination of criteria for stability that do not require knowledge of the general solution of the differential equation.

While the theory of stability is mature with many branches of development, the main results (originally obtained by Poincaré and Lyapunov) follow from two fundamental ideas: linearization and Lyapunov functions.

The Principle of Linearized Stability

A linear first-order differential equation is a differential equation of the form

$$\dot{u} = A(t)u\,, \tag{8}$$

where $A(t)$ denotes an $n \times n$ matrix-valued function of the independent variable t.

Most of the analysis of autonomous linear differential equations can be reduced to linear algebra. Indeed, a function of the form $e^{\lambda t}v$, where λ is a complex number and $v \neq 0$ is a complex vector with n components, is a complex solution of the differential equation

$$\dot{u} = Au\,, \tag{9}$$

if and only if $Av = \lambda v$; that is, λ is an eigenvalue of the matrix A and v is a corresponding eigenvector. The real and imaginary parts of a complex solution are automatically real solutions of the real differential Eq. (9). Also, linear combinations of solutions of linear equations are again solutions. Thus, in case there is a basis of eigenvectors, all solutions of the linear system can be expressed as linear combinations (superpositions) of the special exponential solutions. The general solution, for an arbitrary system matrix A, can always be obtained from linear combinations of solutions of the form $p(t)e^{\lambda t}v$, where $p(t)$ is a polynomial of degree at most $n-1$ and $Av = \lambda v$. In fact, there is always a linear change of variables $w = Bu$, given by some invertible matrix B, such that the new system

$$\dot{w} = Jw\,,$$

where the transformed system matrix $J := BAB^{-1}$ is in Jordan canonical form (see, for example, [78]). This block diagonal system can be solved by linear combinations of solutions of the form $p(t)e^{\lambda t}v$, and the solution of the original system (9) is obtained by the inverse change of variables. Alternatively, it is not difficult to prove that the general solution of the system (9) is given by

$$\phi(t,s,v) = e^{(t-s)A}v\,, \tag{10}$$

where

$$e^{tA} := I + \sum_{k=1}^{\infty}(tA)^k \tag{11}$$

and v is an arbitrary point in $\mathbb{R}^n$.

Solutions of the form $t \mapsto p(t)e^{\lambda t}v$ converge to zero as $t \to \infty$ whenever the real part of the eigenvalue λ is less than zero. This observation leads to a basic result:

Theorem 2 *If all eigenvalues of the matrix A have negative real parts, then the zero solution of the autonomous system (9) is asymptotically stable.*

To examine the stability of a solution ψ of the (perhaps nonlinear) differential Eq. (1), consider a second solution γ, define the deviation $\delta = \gamma - \psi$, and note that

$$\dot{\delta} = f(\gamma,t) - f(\psi,t) = f(\delta+\psi,t) - f(\psi,t)$$

Definition 3 The linearized equation along the solution ψ of the differential Eq. (1) is

$$\dot{w} = f_u(\psi(t), t)w\,, \tag{12}$$

where the subscript denotes the partial derivative with respect to u.

The linearized equation may be viewed as an approximation of the differential equation for the deviation δ because the linearized equation is also obtained by Taylor expanding the function

$$\delta \mapsto f(\delta + \psi, t) - f(\psi, t)$$

to first-order at $\delta = 0$ and ignoring the remainder term. The principle of linearized stability states that the original solution ψ is stable whenever the zero solution of the linearized equation is stable.

The principle of linearized stability is not a theorem; rather, it serves as the underlying idea for the theory of stability by linearization. A basic result of this theory is the following theorem (see [63]).

Theorem 4 *Let*

$$\dot{u} = Au + B(t)u + g(u, t)\,, \quad u(t_0) = u_0\,, \quad u \in \mathbb{R}^n$$

be a smooth initial value problem with solution $t \mapsto \psi(t)$. If

(1) A is a constant matrix and all its eigenvalues have negative real parts,
(2) B is an $n \times n$ matrix valued function, continuously dependent on t, such that the matrix norm $\|B(t)\|$ converges to zero as $t \to \infty$,
(3) g is smooth and there are constants $a > 0$ and $k > 0$ such that $|g(v, t)| \leq k|v|^2$ for all $t \geq 0$ and $|v| > a$,

then there are constants $C > 1$, $\beta > 0$, and $\lambda > 0$ such that

$$|\psi(t)| \leq C|u_0|e^{-\lambda(t-t_0)}\,, \quad t \geq t_0$$

whenever $|u_0| \leq \beta/C$. In particular, the zero solution is asymptotically stable.

Theorem 4 is used for the stability analysis of rest points of autonomous first-order differential equations. If $f\colon \mathbb{R}^n \to \mathbb{R}^n$ and $f(\xi) = 0$, for some $\xi \in \mathbb{R}^n$, then the constant function $\phi(t, \xi) = \xi$ is a solution of the autonomous differential equation

$$\dot{u} = f(u)\,. \tag{13}$$

In this case, ξ is called a rest point for the differential equation. By Taylor expansion at the rest point—this time keeping the remainder term, the differential equation is recast in the form

$$\dot{\gamma} = Df(\xi)\gamma + R(\xi, \gamma)\,,$$

where Df denotes the derivative of f and $\gamma := \phi - \xi$ for an arbitrary solution ϕ. If f is class C^2, we have that

$$|R(\xi, v)| \leq k|v|^2\,.$$

Thus, under the hypothesis that all eigenvalues of the constant system matrix $Df(\xi)$ have negative real parts, Theorem 4 implies that the rest point ξ is asymptotically stable. The hypothesis that f is class C^2 can be relaxed (see, for example, [19]):

Theorem 5 *If $f\colon \mathbb{R}^n \to \mathbb{R}^n$ is class C^1, $\xi \in \mathbb{R}^n$, $f(\xi) = 0$, and all eigenvalues of the matrix $Df(\xi)$ have negative real parts, then the rest point ξ is asymptotically stable. If the matrix $Df(\xi)$ has an eigenvalue with positive real part, then the rest point ξ is unstable.*

The damped harmonic oscillator (5) is equivalent to the first-order system of (linear) differential equations

$$\dot{u} = Au\,, \tag{14}$$

where the system matrix is

$$A := \begin{pmatrix} 0 & 1 \\ -1 & -\epsilon \end{pmatrix}. \tag{15}$$

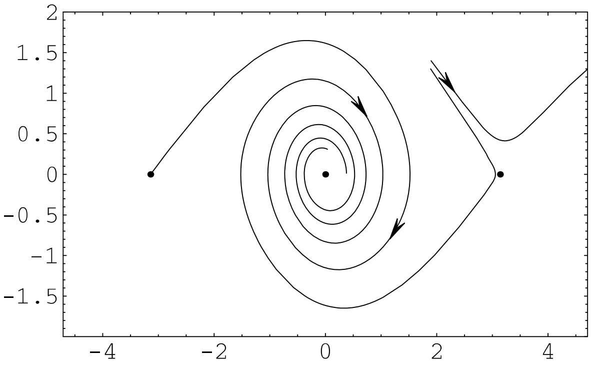
If $\epsilon > 0$, the system matrix has eigenvalues $\frac{1}{2}(-\epsilon \pm \sqrt{\epsilon^2 - 4})$ whose real parts are always negative. Thus, the rest point at the origin is asymptotically stable in agreement with physical intuition.

The power of the linearization method is demonstrated by its applications to systems whose general solutions are not known explicitly; for example, the nonlinear oscillator

$$\ddot{x} + \epsilon x + x + g(x) = 0\,, \tag{16}$$

where g is an arbitrary (smooth) function such that $g(0) = Dg(0) = 0$. In this case, the origin is a rest point of the equivalent first-order system and the system matrix of the linearized system at the origin is again the matrix (15); therefore, by Theorem 5, the origin is asymptotically stable. This result applies to the rest point $(\theta, \dot{\theta}) = (0, 0)$ of the damped pendulum

$$\ddot{\theta} + \epsilon\dot{\theta} + \sin\theta = 0\,. \tag{17}$$



Stability Theory of Ordinary Differential Equations, Figure 4
The figure depicts portions of six orbits in the phase portrait of the damped pendulum (17) with $\epsilon = 0.2$: the rest points at $(\theta, \dot\theta)$ equal to $(-\pi, 0)$, $(0, 0)$, and $(\pi, 0)$; and, the orbits starting at $(-\pi, 0.001)$, $(\pi - 1.25, 1.3)$, and $(\pi - 1.243, 1.4)$. The last two orbits pass close to the (saddle type) unstable rest point at $(\pi, 0)$

The differential Eq. (17) also has a rest point at $(\theta, \dot\theta) = (\pi, 0)$ where the system matrix of the linearized differential equation is

$$\begin{pmatrix} 0 & 1 \\ 1 & -\epsilon \end{pmatrix} . \tag{18}$$

If $\epsilon > 0$, this matrix has an eigenvalue with positive real part; therefore, this rest point—which corresponds to the upward vertical equilibrium of the physical pendulum—is unstable (cf. Fig. 4).

Definition 6 An $n \times n$ matrix is called infinitesimally hyperbolic if all of its eigenvalues have non-zero real parts. A rest point of an autonomous system is called hyperbolic if the system matrix of the linearized equation at this rest point is infinitesimally hyperbolic. The rest point is called nondegenerate if the system matrix is invertible.

An important result on the principle of linearized stability for rest points is the Grobman–Hartman theorem (see [30,31,34,35] and [19,22]):

Theorem 7 *If v is a hyperbolic rest point for the autonomous differential equation $\dot u = f(u)$, then there is an open set V containing v and a homeomorphism H with domain V such that the orbits of this differential equation in V are mapped by H to orbits of the linearized system $\dot w = Df(v)w$.*

In other words, the qualitative behavior of an autonomous differential equation in a sufficiently small neighborhood of a hyperbolic rest point is the same as the behavior of its linearization at this rest point.

The related problem of the existence of a change of variables (a diffeomorphism) taking a given system to its linearization in a neighborhood of a rest point is more difficult to resolve. Fundamental work in this area is due to Poincaré [74], Carl Siegel [88], Henri Dulac [25], Shlomo Sternberg [90,91] and A.D. Brjuno [16] (see also [9,89]). The non-existence of certain relationships (resonances) among the eigenvalues of the system matrix of the linearization plays an important role in the theory.

Definition 8 The vector of eigenvalues $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_n)$ of the $n \times n$-matrix A is called resonant if there is an n-vector of nonnegative integers $M = (m_1, m_2, \dots, m_n)$ such that $\sum_{k=1}^{n} |m_k| \geq 2$ and at least one of the eigenvalues λ_k is given by

$$\lambda_k = \langle M, \lambda \rangle ,$$

where the angle brackets denote the usual inner product. The eigenvalues are nonresonant if no such relationship occurs.

The first theorem of the subject was proved by Poincaré:

Theorem 9 *If the eigenvalues of the linearization of a vector field given by a formal power series are nonresonant, then there is a change of variables in the form of a formal power series that transforms the vector field to its linearization.*

In case the linearization is resonant, there is a formal change of variables that removes all of the terms in the formal series that defines the vector field except those that are resonant (that is, the corresponding monomials are given by products of variables whose powers form vectors with the properties of M in the definition of resonance).

Definition 10 The vector of eigenvalues $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_n)$ of the $n \times n$-matrix A belongs to the Poincaré domain if the convex hull of its components in the complex plane does not contain the origin; otherwise, the vector belongs to the Siegel domain. The vector λ is called type (C, ν) for $C > 0$ and $\nu > 0$, if the inequality

$$|\lambda_k - \langle M, \lambda \rangle| \geq \frac{C}{(\sum_k |m_k|)^\nu}$$

is satisfied for each component λ_k of λ and all vectors M as in Definition 8.

Theorem 11 *If the eigenvalues of the linearization of an analytic vector field are nonresonant and in the Poincaré domain, then there is an analytic change of coordinates that transforms the vector field to its linearization. If the eigenvalues of the linearization of an analytic vector field are of* *type (C, ν), then there is an analytic change of coordinates that transforms the vector field to its linearization.*

For modern proofs of Poincaré's and Siegel's theorems see [71] or [87]. Sternberg proved Siegel's result for vector fields in the Siegel domain for sufficiently smooth vector fields (see [90,91]).

To successfully apply linearized stability for rest points of autonomous systems, the fundamental problem is to determine the real parts of the eigenvalues of a square matrix; especially, it is important to determine conditions that ensure all eigenvalues have negative real parts. The most important result in this direction is the Routh-Hurwitz criterion (see [40,84]):

Theorem 12 *Suppose that the characteristic polynomial of the real matrix A is written in the form*

$$\lambda^n + a_1\lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n \, ,$$

let $a_m = 0$ for $m > n$, and define the determinants Δ_k for $k = 1, 2, \ldots, n$ by

$$\Delta_k := \det \begin{pmatrix} a_1 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ a_3 & a_2 & a_1 & 1 & 0 & 0 & \cdots & 0 \\ a_5 & a_4 & a_3 & a_2 & a_1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{2k-1} & a_{2k-2} & a_{2k-3} & a_{2k-4} & a_{2k-5} & a_{2k-6} & \cdots & a_k \end{pmatrix} .$$

If $\Delta_k > 0$ for $k = 1, 2, \ldots n$, then all roots of the characteristic polynomial have negative real parts.

The principle of linearized stability is not valid in general. In particular, the stability of the linearized equation at a rest point of an autonomous differential equation does not always determine the stability of the rest point. For example, consider the scalar differential equation $\dot{u} = -u^m$, where $m > 1$ is a positive integer. The origin $u = 0$ is a rest point for this differential equation and the corresponding linearized system is $\dot{w} = 0$. The origin (and every other point on the real line) is a stable rest point of this linear differential equation. On the other hand, by solving the original equation, it is easy to see that the origin is asymptotically stable in case m is odd and unstable in case m is even. Thus, the stability of the linearization does not determine the stability of the rest point.

There is no theorem known that can be used to determine the stability of a general nonautonomous first-order linear differential Eq. (8). In particular, there is no obvious relation between the (time-dependent) eigenvalues of a time-dependent matrix and the stability of the zero solution of the corresponding first-order linear differential equation. Consider, as an example, the π-periodic system $\dot{u} = A(t)u$, where

$$A(t) = \begin{pmatrix} -1 + \frac{3}{2}\cos^2 t & 1 - \frac{3}{2}\sin t \cos t \\ -1 - \frac{3}{2}\sin t \cos t & -1 + \frac{3}{2}\sin^2 t \end{pmatrix} , \quad (19)$$

which was constructed by Lawrence Marcus and Hidehiko Yamabe [66]. The real parts of the eigenvalues of $A(t)$ (given by $\frac{1}{4}(-1 \pm \sqrt{7}\, i)$) are negative; but,

$$u(t) = e^{t/2} \begin{pmatrix} -\cos t \\ \sin t \end{pmatrix}$$

is a solution that grows without bound.

While a general theory of linearized stability for non-constant solutions of differential equations remains an area of research, much is known in case the non-constant solution is periodic.

Suppose that $t \mapsto \psi(t)$ is a periodic solution with period T of the differential Eq. (1) defined on $\mathbb{R}^n$. Linearization leads to the differential equation

$$\dot{\gamma} = f_u(\psi(t), t)\gamma \, , \quad (20)$$

where the matrix (valued function) $A(t) := f_u(\psi(t), t)$ is periodic with period T. Thus, the theory of linearized stability for periodic solutions starts with the stability theory for the periodic linear differential Eq. (8). The most basic results in this setting are due to Gaston Floquet [27]. To state them, let us first recall that a linear differential equation admits matrix solutions; that is, matrices whose columns are vector solutions. A fundamental matrix solution is an $n \times n$-matrix solution with linearly independent columns; or, equivalently, a matrix solution whose values are all invertible as $n \times n$-matrices.

Floquet's main results are stated in the following theorem (see [19] and recall the definition of the matrix exponential in display (11)).

Theorem 13 *If $\Phi(t)$ is a fundamental matrix solution of the T-periodic system (8), then*

$$\Phi(t + T) = \Phi(t)\Phi^{-1}(0)\Phi(T)$$

for all $t \in \mathbb{R}$. In addition, there is a matrix B (which may be complex) such that

$$e^{TB} = \Phi^{-1}(0)\Phi(T)$$

and a T-periodic matrix function $t \mapsto P(t)$ (which may be complex valued) such that $\Phi(t) = P(t)e^{tB}$ for all $t \in \mathbb{R}$. Also, there is a real matrix R and a real $2T$-periodic matrix function $t \to Q(t)$ such that $\Phi(t) = Q(t)e^{tR}$ for all $t \in \mathbb{R}$.

The stability of the zero solution of the periodic linear differential equation is intimately connected with the eigenvalues of the matrix R in Floquet's theorem. Indeed, the time-dependent (real) change of variables $u = Q(t)v$ transforms the T-periodic differential Eq. (8) to the autonomous linear system

$$\dot{v} = Rv \, .$$

Definition 14 The representation $\Phi(t) = P(t)\mathrm{e}^{tB}$ of the fundamental matrix solution Φ in Floquet's theorem is called a Floquet normal form. The eigenvalues of the matrix e^{tB} are called the characteristic (or Floquet) multipliers of the corresponding linear system. A complex number μ is called a characteristic (or Floquet) exponent if there is a characteristic multiplier ρ such that $\mathrm{e}^{\mu T} = \rho$.

Theorem 15

(1) *The characteristic multipliers and the characteristic exponents do not depend on the choice of the fundamental matrix solution of the T-periodic differential Eq. (8).*
(2) *If the characteristic multipliers of the periodic system (8) all have modulus less than one; equivalently, if all characteristic exponents have negative real parts, then the zero solution is asymptotically stable.*
(3) *If the characteristic multipliers of the periodic system (8) all have modulus less than or equal to one; equivalently, if all characteristic exponents have nonpositive real parts, and if the algebraic multiplicity equals the geometric multiplicity of each characteristic multiplier with modulus one; equivalently, if the algebraic multiplicity equals the geometric multiplicity of each characteristic exponent with real part zero, then the zero solution is stable.*
(4) *If at least one characteristic multiplier of the periodic system (8) has modulus greater than one; equivalently, if a characteristic exponent has positive real part, then the zero solution is unstable.*

While Theorem 15 gives a complete description of the stability of the zero solution, no general method is known to determine the characteristic multipliers (cf. [98]). The power of the theorem is an immediate corollary: A finite set of numbers (namely, the characteristic multipliers) determine the stability of the zero solution, at least in the case where none of them have modulus one. The characteristic multipliers can be approximated by numerical integration. And, in special cases, their moduli can be determined by mathematical analysis.

One of the most important applications for Floquet theory and the method of linearization, is the analysis of Hill's equation

$$\ddot{x} + a(t)x = 0 \, , \tag{21}$$

or, equivalently, the first-order linear system

$$\dot{x} = y \, , \quad \dot{y} = -a(t)x \, , \tag{22}$$

where x is a scalar and a is a T-periodic function (see [19, 64]). This equation arose from George Hill's study of lunar motion. It is ubiquitous in stability theory.

The stability of the zero solution of Hill's system (22) can be reduced to the identification of a single number; namely, the magnitude of the trace of the principal fundamental matrix solution $\Phi(t)$ at $t = T$ (that is, the fundamental matrix solution $\Phi(t)$ such that $\Phi(0) = I$, where I is the $n \times n$ identity matrix).

Theorem 16 *Suppose that $\Phi(t)$ is the principal fundamental matrix solution of system (22). If $|\operatorname{trace} \Phi(T)| < 2$, then the zero solution is stable. If $|\operatorname{trace} \Phi(T)| > 2$, then the zero solution is unstable.*

This result, of course, requires knowledge of the solutions of the system. On the other hand, a beautiful theorem of Lyapunov [55] gives a stability criterion using only properties of the function a:

Theorem 17 *If $a \colon \mathbb{R} \to \mathbb{R}$ is a positive T-periodic function such that*

$$T \int_0^T a(t)\,dt \le 4 \, ,$$

then all solutions of the Hill's equation $\ddot{x} + a(t)x = 0$ are bounded. In particular, the trivial solution is stable.

The principle of linearized stability for periodic solutions of nonlinear systems motivates a theory similar to the stability theory for rest points. There is, however, one essential difference that will be explained for a T-periodic solution γ of the autonomous first-order differential equation

$$\dot{u} = f(u). \tag{23}$$

The linearized equation at γ is $\dot{w} = Df(\gamma(t))w$. Let $\phi(t) = f(\gamma(t))$ and note that

$$\dot{\phi}(t) = Df(\gamma(t))f(\gamma(t)) = Df(\gamma(t))\phi(t) \, ;$$

that is, $t \mapsto f(\gamma(t))$ is a solution of the linearized equation. This solution is given by

$$f(\gamma(t)) = \Phi(t)f(\gamma(0)) \, ,$$

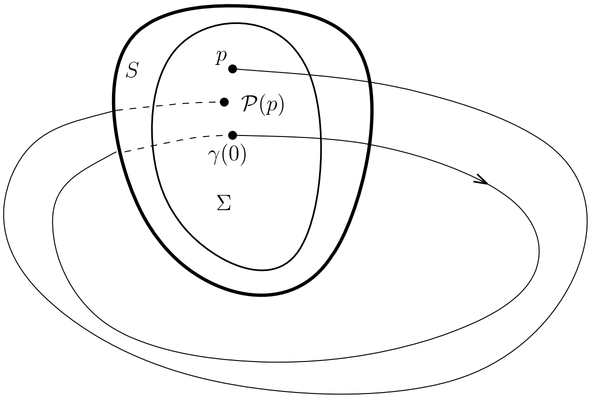
where $\Phi(t)$ is the principal fundamental matrix solution of the linearized equation. If follows that $\Phi(T)f(\gamma(0)) =$

$f(\gamma(0))$; therefore, the number 1 is a characteristic multiplier. It corresponds to motion in the direction of the periodic solution and does not contribute to the stability or instability of the periodic solution. This characteristic multiplier must be treated separately in the stability theory.

Theorem 18 *If the number 1 occurs with algebraic multiplicity one in the set of characteristic multipliers of the linearized differential equation at a periodic solution and all other characteristic multipliers have modulus less than one, then the periodic solution is orbitally asymptotically stable. If at least one characteristic multiplier has modulus larger than one, then the periodic solution is unstable.*

Poincaré introduced a geometric approach to stability for periodic solutions. His idea is simple and useful: Consider a periodic solution γ and some point, say $\gamma(0)$, on the orbit of this solution in $\mathbb{R}^n$. Construct a hypersurface S in $\mathbb{R}^n$ that contains $\gamma(0)$ and is transverse to the orbit of γ at this point. By the implicit function theorem, there is an open neighborhood Σ of $\gamma(0)$ in S such that the solution of the differential equation starting at every point in Σ returns to S. This defines a local diffeomorphism $\mathcal{P}\colon \Sigma \to S$, called the Poincaré (or return) map, which assigns p to its point of first return on S (cf. Fig. 5).

The eigenvalues of the derivative of $\mathcal{P}$ at $\gamma(0)$ are exactly the characteristic multipliers associated with γ. Thus, if all eigenvalues of this derivative have modulus less than one, then γ is orbitally asymptotically stable. The derivative of the Poincaré map can be computed using the linearized equations of the corresponding differential equation along its the periodic orbit. While explicit formulas for the desired derivative are only available for special cases, the Poincaré map provides an indispensable tool for organizing the study of stability for periodic orbits in general.



Stability Theory of Ordinary Differential Equations, Figure 5
A schematic diagram of a Poincaré map $\mathcal{P}$ defined on a Poincaré section Σ near a periodic orbit containing the point $\gamma(0)$

It should be clear that the discrete dynamical system defined by $\mathcal{P}$ codes all the information about the qualitative behavior of the solutions of the differential equation that start near the orbit of γ. The existence of the Poincaré map is thus the main connection between discrete and continuous dynamical systems.

Often, but not always, the dynamical properties of diffeomorphisms (or maps) are simpler to analyze than the corresponding properties of differential equations. Thus, theorems are often proved first for maps and later for differential equations. For instance, the Grobman–Hartman theorem was first proved for diffeomorphisms. This result can be applied to the Poincaré map of a periodic orbit to prove the existence of a homeomorphism that maps nearby orbits to the orbits of a corresponding linearization.

For the special case of periodic solutions of autonomous first-order differential equations on the plane, there is exactly one important Floquet multiplier; it is identified in the following fundamental result.

Theorem 19 *If γ is a periodic solution with period T of the differential Eq. (23) defined on $\mathbb{R}^2$, then the number*

$$\mu = \int_0^T \operatorname{div} f(\gamma(t))\,dt$$

(where div *denotes the divergence) is (up to a positive scalar multiple) a characteristic exponent of γ. If μ is negative, then γ is orbitally asymptotically stable. If μ is positive, then γ is unstable.*

The system of differential equations

$$\dot{x} = x - y - (x^2 + y^2)x\,, \quad \dot{y} = x + y - (x^2 + y^2)y$$

has the periodic solution $t \to (\cos t, \sin t)$ (depicted with a non-unit aspect ratio in Fig. 2). Also, the divergence of the vector field is $2 - 4(x^2 + y^2)$; it has value -2 along the corresponding periodic orbit $\Gamma = \{(x, y)\colon x^2 + y^2 = 1\}$. By Theorem 19, Γ is orbitally asymptotically stable. This is an example of a limit cycle; that is, an isolated periodic orbit. By changing to polar coordinates (r, θ), the transformed system

$$\dot{r} = r(1 - r^2)\,, \quad \dot{\theta} = 1$$

decouples, and its flow is given by

$$\phi_t(r, \theta) = \left(\left(\frac{r^2 e^{2t}}{1 - r^2 + r^2 e^{2t}} \right)^{\frac{1}{2}}, \theta + t \right). \tag{24}$$

The positive x-axis is a Poincaré section with corresponding Poincaré map P given by

$$P(x) = \left(\frac{x^2 e^{4\pi}}{1 - x^2 + x^2 e^{4\pi}} \right)^{\frac{1}{2}} .$$

The periodic orbit Γ intersects the Poincaré section at $x = 1$ where we have $P(1) = 1$ and $P'(1) = e^{-4\pi} < 1$.

A periodic solution of an autonomous system cannot be asymptotically stable. This fact depends on a special property of the solutions of autonomous systems: if the differential Eq. (1) is autonomous, then the function ψ giving the solutions, $t \mapsto \psi(t, s, v)$, is a function of $t - s$. Hence, by replacing $t - s$ by t, each solution can be expressed in the form $t \mapsto \phi(t, v)$ where $\phi(0, v) = v$ and $\phi(t, \phi(s, v)) = \phi(t + s, v)$ whenever both sides are defined. The family of functions $v \mapsto \phi(t, v)$, parametrized by t is called the flow of the differential Eq. (1).

Suppose that $t \mapsto \phi(t, v)$ is an asymptotically stable periodic solution with period T. Let $\delta > 0$ be as in Definition 1. If $|w - v| < \delta$, then $\lim_{t\to\infty} |\phi(t, w) - \phi(t, v)| = 0$. For all sufficiently small s, we have that $|\phi(s, w) - v| < \delta$. Hence, $\lim_{t\to\infty} |\phi(t, \phi(s, w)) - \phi(t, v)| = 0$. By the triangle law,

$$\begin{aligned} &|\phi(t + s, v) - \phi(t, v)| \\ &\le |\phi(t + s, v) - \phi(t + s, w)| + |\phi(t + s, w) - \phi(t, v)| \\ &\le |\phi(s, \phi(t, v)) - \phi(s, \phi(t, w)| + |\phi(t + s, w) - \phi(t, v)|. \end{aligned}$$

Since $\xi \mapsto \phi(s, \xi)$ is smooth and a periodic orbit is compact, the law of the mean implies that there is a constant $C > 0$ such that

$$\begin{aligned} &|\phi(t + s, v) - \phi(t, v)| \\ &\quad \le C|\phi(t, v) - \phi(t, w)| + |\phi(t + s, w) - \phi(t, v)| . \end{aligned}$$

Both summands on the right-hand side of the last inequality converge to zero as $t \to \infty$. Hence, we have that

$$\lim_{k\to\infty} |\phi(kT + s, v) - \phi(kT, v)| = 0 ,$$

where k denotes an integer variable. It follows that $\phi(s, v) = v$ for all s in some open set. The point v must be a rest point on a periodic orbit, in contradiction to the uniqueness of solutions.

While a periodic solution of an autonomous system cannot be asymptotically stable, we can ask that an orbitally asymptotically stable periodic orbit satisfy a stronger type of stability.

Definition 20 Let ϕ denote the flow of the autonomous differential Eq. (23) and suppose that Γ is a periodic orbit corresponding to the solution $t \mapsto \phi(t, p)$. We say that $q \in \mathbb{R}^n$ has asymptotic phase p with respect to Γ if the solution $t \mapsto \phi(t, q)$ is such that

$$\lim_{t\to\infty} |\phi(t, q) - \phi(t, p)| = 0 .$$

We say that Γ is isochronous if there exists an open neighborhood of its orbit such that every point in this neighborhood has asymptotic phase with respect to the period solution.

Theorem 21 *If all eigenvalues of the derivative of a Poincaré map at the periodic solution $t \mapsto \phi(t, p)$ of (23) lie strictly inside the unit circle in the complex plane (equivalently, the corresponding periodic orbit is hyperbolic), then every point q in some open neighborhood of this periodic orbit has asymptotic phase.*

A theory of asymptotic phase that includes orbitally stable periodic orbits such that some eigenvalues of the derivative of an associated Poincaré map have modulus one was recently completed (see [21,26]).

Lyapunov Functions

In case the linearized differential equation along a solution is stable but not orbitally asymptotically stable, the principle of linearized stability usually cannot be justified. One of the great contributions of Lyapunov is the introduction of a method that can be used to determine the stability of such solutions.

The fundamental ideas of Lyapunov are most clear for the stability analysis of rest points of autonomous systems (see [63]); but, the theory goes far beyond this case (see [51]).

Definition 22 Let u_0 be a rest point of the autonomous differential Eq. (23). A continuous function $L\colon U \to \mathbb{R}$, where $U \subseteq \mathbb{R}^n$ is an open set with $u_0 \in U$, is called a Lyapunov function at u_0 if

(1) $L(u_0) = 0$,
(2) $L(x) > 0$ for $x \in U \setminus \{u_0\}$,
(3) the function L is continuously differentiable on the set $U \setminus \{u_0\}$, and, on this set, $\dot{L}(x) := \operatorname{grad} L(x)\cdot f(x) \le 0$.

The function L is called a strict Lyapunov function if, in addition,

(4) $\dot{L}(x) < 0$ for $x \in U \setminus \{u_0\}$.

Property (3) states that L does not increase along solutions.

Theorem 23 *If there is a Lyapunov function defined on an open neighborhood of a rest point of an autonomous first-order differential equation, then the rest point is stable. If, in* *addition, the Lyapunov function is a strict Lyapunov function, then the rest point is asymptotically stable.*

Lyapunov's theorem was motivated by Lagrange's principle: a rest point of a mechanical system corresponding to an isolated minimum of the potential energy is stable. To obtain this result, recall that the equation of motion of a (conservative) mechanical system for the motion of a particle is

$$m\ddot{x} = -\operatorname{grad} G(x)$$

where m is the mass of the particle and $x \in \mathbb{R}^n$ is its position. The kinetic energy of the particle is defined to be $K = \frac{m}{2}\langle \dot{x}, \dot{x}\rangle$ (where the angle brackets denote the usual inner product) and the potential energy is G, a quantity defined up to an additive constant.

Consider the corresponding equivalent first-order system

$$\dot{x} = y\,, \quad \dot{y} = -\frac{1}{m}\operatorname{grad} G(x)\,,$$

and suppose that x_0 be an isolated minimum of G. The state $(x, y) = (x_0, 0)$ is a rest point of the mechanical system. By inspection, the total energy

$$L(x, y) := \frac{m}{2}\langle y, y\rangle + G(x) - G(x_0)$$

(with an appropriate translation of the potential energy) is a Lyapunov function at the rest point. Thus, the rest point is stable.

The proof of the first part of Lyapunov's theorem is simple: Choose an open ball B with center at the rest point and radius sufficiently small so that its closure is in the domain of the Lyapunov function L. The continuous positive function L on the closed and bounded spherical boundary of B has a non-zero minimum value b. Because the continuous function L vanishes at the rest point, there is a second ball A that is concentric with B, has strictly smaller radius, and is such that the maximum value of L on all of A is smaller than b. A solution of the differential equation starting in A cannot reach the boundary of B because L does not increase along solutions. Thus, the rest point is stable.

Lyapunov's method extends to nonautonomous systems with an appropriate modification of the notion of a Lyapunov function. Suppose that the first-order differential Eq. (1) has a rest point in the sense that for some $T \in \mathbb{R}$, $f(u_0, t) = 0$ for all $t \geq T$. A Lyapunov function is a class C^1 function defined in a neighborhood of u_0 for all $t \geq T$ such that $L(u_0, t) = 0$ for $t \geq T$, there is a continuous nonnegative function M defined on the same neighborhood of u_0 such that $L(u, t) \geq M(u)$ for all $t \geq T$ and $\dot{L} = L_t + L_u \cdot \dot{u} \leq 0$ for all $t \geq T$. If such a Lyapunov function exists, then the rest point is stable. If, in addition, there is a positive function N with the same domain as M such that $\dot{L}(u, t) \leq -N(u)$ for all $t \geq T$, then the rest point is asymptotically stable.

Lyapunov's direct method can be used to prove the principle of linearized stability for rest points of autonomous systems. To see how this might be done, consider the differential equation

$$\dot{u} = Au + g(u)\,, \quad u \in \mathbb{R}^n\,,$$

where A is a real $n \times n$ matrix and $g\colon \mathbb{R}^n \to \mathbb{R}^n$ is a smooth function. Suppose that every eigenvalue of A has negative real part, and that for some $a > 0$, there is a constant $k > 0$ such that, using the usual norm on $\mathbb{R}^n$,

$$|g(x)| \leq k|x|^2$$

whenever $|x| < a$. Let $\langle\cdot,\cdot\rangle$ denote the usual inner product on $\mathbb{R}^n$, and let A^* denote the transpose of the real matrix A. Suppose that there is a real symmetric positive definite $n \times n$ matrix that also satisfies Lyapunov's equation

$$A^*B + BA = -I$$

and define $L\colon \mathbb{R}^n \to \mathbb{R}$ by

$$L(x) = \langle x, Bx\rangle\,.$$

Using Schwarz's inequality, it can be proved that the restriction of L to a sufficiently small neighborhood of the origin is a strict Lyapunov function. The proof is completed by showing that there is a symmetric positive-definite solution of Lyapunov's equation. In fact,

$$B := \int_0^\infty \mathrm{e}^{tA^*}\mathrm{e}^{tA}\,\mathrm{d}t$$

is a symmetric positive definite $n \times n$ matrix that satisfies Lyapunov's equation. Alternatively, it is possible to prove the existence of a solution to Lyapunov's equation using purely algebraic methods (see, for example, [97]).

The instability of solutions can also be detected by Lyapunov's methods. A simple result in this direction is the content of the following theorem.

Theorem 24 *Suppose that L is a smooth function defined on an open neighborhood U of the rest point u_0 of the autonomous differential Eq. (23) such that $L(u_0) = 0$ and $\dot{V}(u) > 0$ on $U \setminus \{u_0\}$. If L has a positive value somewhere in each open set containing u_0, then u_0 is unstable.*

One indication of the subtlety of the determination of stability is the insolvability of the center-focus problem [41].

Consider a planar system of first-order differential equations in the form

$$\dot{x} = y + P(x, y)\,, \quad \dot{y} = -x + Q(x, y)\,,$$

where P and Q are analytic at the origin with leading-order terms at least quadratic in x and y. Is the origin a focus (that is, asymptotically stable or asymptotically unstable) or a center (that is, all orbits in some neighborhood of the origin are periodic)? There is no algorithm that can be used to solve this problem for all such systems in a finite number of steps. On the other hand, the center-focus problem is solved, by using Lyapunov's stability theorem, if a certain sequence of numbers called Lyapunov quantities (which can be computed iteratively and algebraically) has a non-zero element. One problem is that the algorithm for computing the Lyapunov quantities may not terminate in a finite number of steps (see, for example, [19,89]).

The center-focus problem is solved for some special cases. The most important theorem in this area of research is due to Nikolai Bautin (see [14] and [99,100] for a modern proof); he found (among other results) a complete solution of the center-focus problem for quadratic systems (that is, where P and Q are homogeneous quadratic polynomials). While the center-focus problem for quadratics had been solved by Dulac [25] and others [28,43,44,45] before his paper appeared; Bautin introduced a method, which involves the use of polynomial ideals, that has had far reaching consequences.

Stability in Conservative Systems and the KAM Theorem

Stability theory for conservative systems leads to many delicate problems, some of which—for example Lagrange's problem—can be resolved by the method of Lyapunov functions. While the total energy, the total angular momentum, or other first integrals all have the property that their derivatives vanish along trajectories, these integrals do not always serve as Lyapunov functions because they often fail to be positive definite in punctured neighborhoods of rest points or periodic orbits. Thus, other methods are required.

As in the quest to determine the stability of the solar system, a basic problem in mechanics is to determine the stability of periodic motions (or rest points) of conservative differential equations with respect to small perturbations. While no general solution is known, great progress has been made culminating in the Kolmogorov–Arnold–Moser (KAM) theorem (see [5,46,47,68,87,89,92]).

A mechanical system with N degrees-of-freedom (that is, the positions are determined by N-coordinates) is called completely integrable if it has N independent first integrals whose Poisson brackets are in involution (see p. 271 in [8]). In this case, the system can be transformed to a first-order system of differential equations in the simple form

$$\dot{I} = 0, \quad \dot{\theta} = \Omega(I)\,, \tag{25}$$

where I is an N-dimensional variable (of actions), θ is an N-dimensional variable (of angles), and Ω is a smooth function. The new coordinates are called action-angle variables (see, for generalizations, [39,42]). This system is in Hamiltonian form ($\dot{I} = \partial H/\partial\theta$, $\dot{\theta} = -\partial H/\partial I$) with Hamiltonian $H(I, \theta) := -\omega(I)$, where ω is an anti-derivative of Ω; and, of course, the Hamiltonian is constant along solutions of the system.

The corresponding perturbed system is taken to have the form

$$\dot{I} = \epsilon \frac{\partial F}{\partial \theta}(I, \theta), \quad \dot{\theta} = \Omega(I) - \epsilon \frac{\partial F}{\partial I}(I, \theta) \tag{26}$$

(or, more generally, a similar form with additional terms that are higher-order in ϵ), where F is a smooth function that is 2π-periodic in θ. This system is in Hamiltonian form with Hamiltonian

$$\mathcal{H}(I, \theta, \epsilon) := -\omega(I) + \epsilon F(I, \theta)\,.$$

The unperturbed integrable system can be solved explicitly as

$$I(t) = I_0\,, \quad \theta(t) = \Omega(I_0)t + \theta_0;$$

hence, by inspection, all orbits are periodic or quasi-periodic (when the angular variables are defined modulo 2π) according to whether or not the vector of frequencies $\Omega(I_0)$ satisfies a resonance relation, that is, $\langle M, \Omega(I_0)\rangle = 0$ for some N-vector M with integer components.

Every orbit of the unperturbed system with one degree-of-freedom is periodic; but, for example for two degrees-of-freedom, orbits are periodic with period T only if there are integers m_1 and m_2 such that

$$T\Omega_1(I_0) = 2\pi m_1, \quad T\Omega_2(I_0) = 2\pi m_2$$

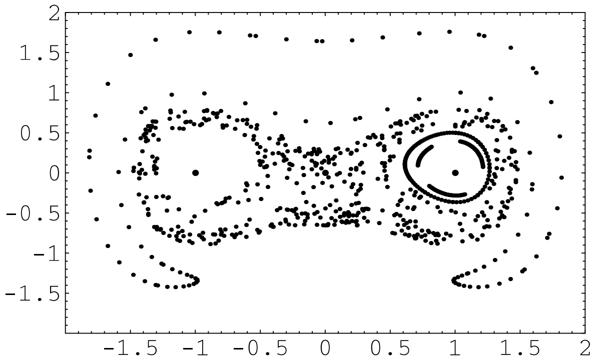
or $m_1\Omega_2(I_0) - m_2\Omega_1(I_0) = 0$.

Geometrically, a choice of actions fixes an energy level and the angles correspond to an N-dimensional invariant torus. The flow on this torus is either periodic or quasi-periodic according to the existence of a corresponding resonance relation. For this reason, the tori with periodic flows are called resonant; the others are called nonresonant. The KAM theorem states that most of the nonresonant tori in the unperturbed system (25) survive after small perturbations as invariant tori of the perturbed system (26).

Theorem 25 *If (in the Hamiltonian system (26)) $\mathcal{H}$ is sufficiently smooth, $D\Omega(I)$ is invertible, and ϵ is sufficiently small, then almost all (in the sense of Lebesgue measure) nonresonant unperturbed invariant tori persist.*

There are similar theorems for time-dependent perturbations of integrable systems and for area-preserving maps (see [8]).

For a mechanical system with N degrees-of-freedom, the perturbed invariant tori are N-dimensional manifolds embedded in the $2N$-dimensional phase space. After restriction to a regular energy surface (that is, a level set of $\mathcal{H}$ corresponding to one of its regular values), these N-dimensional tori are embedded in a $(2N-1)$-dimensional manifold. For $N \le 2$, the invariant tori that exist by the KAM theorem separate the energy hypersurface into two components, one inside and one outside of each invariant torus. A motion starting in a bounded region whose boundary is one of these invariant tori cannot escape to the corresponding unbounded component. Since the invariant tori are dense, such motions are stable (cf. Fig. 6). More generally, each perturbed invariant torus is stable as an invariant set. This result can be used, for example, to prove the stability of certain periodic orbits in the restricted three-body problem (see [54]). On the other hand, for $N > 2$, the invariant tori no longer separate energy surfaces into bounded and unbounded components. Thus, it is possible for orbits to migrate outside nearby tori in a conjectured process called Arnold diffusion (see [6]). The validity of Arnold diffusion for the general Hamiltonian system (26) is an area of current research (see, for a review, [60]).



Stability Theory of Ordinary Differential Equations, Figure 6
The figure depicts portions of seven orbits of the (stroboscopic) Poincaré map $\xi \in \mathbb{R}^2 \mapsto \phi(2\pi, \xi)$ for the forced oscillator $\dot{x} = y, \dot{y} = x - x^3 + 0.05 \sin t$: three fixed points at (x, y) equal to $(-1, 0)$, $(0, 0)$, and $(1, 0)$; an orbit starting at $(x, y) = (1.0, 0.4)$ at time $t = 0$, which is close to a $(3 : 1)$ resonance; an orbit starting at $(1.0, 0.5)$, which is on an invariant torus that surrounds the resonant orbit; an orbit starting at $(1.0, 0.6)$, which appears to be "chaotic"; and an orbit starting at $(1, 1)$, which is on a large invariant torus. The $(3 : 1)$ resonant orbit is stable

Averaging and the Stability of Perturbed Periodic Orbits

As we have seen, systems of the form

$$\dot{I} = \epsilon P(I, \theta)\,, \qquad \dot{\theta} = \Omega(I) + \epsilon Q(I, \theta)\,, \tag{27}$$

where P and Q are 2π-periodic, often arise in mechanics. Here, the properties of system (27) are considered for general perturbations, which are not necessarily conservative.

Note that, in case ϵ is small, the vector of actions (components of I) in the differential Eq. (27) evolves slowly relative to the evolution of the vector of angles θ. The averaging principle states that the evolution of the actions for such a system is well-approximated by the corresponding averaged equation given by

$$\dot{\mathcal{I}} = \epsilon \bar{P}(\mathcal{I})\,, \tag{28}$$

where

$$\bar{P}(\mathcal{I}) := \frac{1}{(2\pi)^N} \int_{\mathbb{T}^N} P(\mathcal{I}, \theta)\, \mathrm{d}\theta$$

and $\mathbb{T}^N$ denotes the N-dimensional torus of angles.

Exactly what is meant by "well-approximated" is clarified by the averaging theorem, which requires a severe restriction: the vector of angles is one-dimensional (cf. [61,86]).

Theorem 26 *If θ is a scalar variable and Ω is bounded away from zero, then there is a near-identity change of variables of the form $\mathcal{I} = I + \epsilon k(I, \theta)$ that is 2π-periodic in θ which transforms system (27) into the form*

$$\dot{\mathcal{I}} = \epsilon \bar{P}(\mathcal{I}) + \epsilon^2 P_1(\mathcal{I}, \theta, \epsilon)\,, \qquad \dot{\theta} = \Omega(\mathcal{I}) + \epsilon Q_1(\mathcal{I}, \theta, \epsilon)\,.$$

where P_1 and P_2 are 2π-periodic in θ. Moreover, if the evolution of I starting at I_0 is given by $I(t)$ and the evolution (according to the differential Eq. (28)) of the averaged action $\mathcal{I}$ starting at I_0 is given by $\mathcal{I}(t)$, then there are constants $C > 0$ and $T(I_0) > 0$ such that

$$|\mathcal{I}(t) - I(t)| \le C\epsilon$$

on the time interval $0 \le \epsilon t < T(I_0)$.

The existence and stability of perturbed period solutions is the content of the following theorem.

Theorem 27 *Consider the system*

$$\dot{I} = \epsilon F(I,\theta) + \epsilon^2 F_2(I,\theta,\epsilon)\,, \qquad \dot{\theta} = \Omega(I) + \epsilon G(I,\theta,\epsilon)\,, \tag{29}$$

where $I \in \mathbb{R}^M$, θ is a scalar, F, F_2, and G are 2π-periodic functions of θ, and there is some number c such that $\Omega(I) \geq c > 0$. If the averaged system has a nondegenerate rest point (see Definition 6) and ϵ is sufficiently small, then system (29) has a periodic orbit. If in addition $\epsilon > 0$ and the rest point is hyperbolic, then the periodic orbit has the same stability type as the hyperbolic rest point; that is, the dimensions of the corresponding stable and unstable manifolds are the same (see Definition 29).

For a typical application of Theorem 27, consider the system of differential equations

$$\begin{aligned} \dot{I} &= \epsilon(a + b\sin k\theta)\sin\psi + \epsilon^2 P(I,\psi,\theta),\\ \dot{\psi} &= \epsilon c I + \epsilon^2 Q(I,\psi,\theta),\\ \dot{\theta} &= 1, \end{aligned} \tag{30}$$

where a, b, and c are non-zero constants, k is a positive integer, and both P and Q are 2π-periodic in θ. The averaged system of differential equations

$$\dot{I} = \epsilon a \sin\psi\ , \quad \dot{\psi} = \epsilon c I$$

has hyperbolic rest points at $(I,\psi) = (0,0)$ and $(I,\psi) = (0,\pi)$. According to Theorem 27, if ϵ is sufficiently small, the system of differential Eq. (30) has corresponding hyperbolic periodic solutions. In case $ac > 0$, the solution of the perturbed system corresponding to $\psi = 0$ is stable, and the solution corresponding to $\psi = \pi$ is unstable (in fact, it is a hyperbolic saddle). If $ac < 0$, the stability types are switched.

Structural Stability

An important aspect of the global theory of dynamical systems is the stability of the orbit structure as a whole. The motivation for the corresponding theory comes from applied mathematics. Mathematical models always contain simplifying assumptions. Dominant features are modeled; supposed small disturbing forces are ignored. Thus, it is natural to ask if the qualitative structure of the set of solutions—the phase portrait—of a model would remain the same if small perturbations were included in the model. The corresponding mathematical theory is called structural stability.

The correct mathematical formulation of the definition of structural stability requires the introduction of a topology on the space of dynamical systems under consideration. The natural setting for the theory is the space of class C^1 vector fields on a smooth compact manifold.

There is a one-to-one correspondence between vector fields and differential equations: the right-hand side of a first-order autonomous differential equation defines a vector field on $\mathbb{R}^n$. The local representatives, with respect to coordinate charts, of a vector field on an n-dimensional manifold M are vector fields on $\mathbb{R}^n$. The set of all smooth vector fields $\mathfrak{X}(M)$ on M has the structure of a Banach space after the introduction of a Riemannian metric. Thus, two vector fields are close if the distance between them in this Banach space is small.

Definition 28 A vector field X on a smooth manifold M is called structurally stable, if there is some neighborhood U of X in $\mathfrak{X}(M)$ such that for each $Y \in U$ there is a homeomorphism of M that maps orbits of Y to orbits of X and preserves the orientation of orbits according to the direction of time. Such a homeomorphism is called a topological equivalence.

In other words, all perturbations of a structurally stable vector field have the same qualitative orbit structures.

At first impression, it might seem that the homeomorphism in the definition of structural stability should be a diffeomorphism. But, this requirement is too strong. To see why, consider a hyperbolic rest point u_0 of the first-order autonomous differential Eq. (23) in $\mathbb{R}^n$. Let $\dot{u} = g(u)$ be a differential equation on $\mathbb{R}^n$, and consider the perturbation of the differential Eq. (23) given by $\dot{u} = f(u) + \epsilon g(u)$. We have that $f(u_0) = 0$. Thus, $F(u,\epsilon) := f(u) + \epsilon g(u)$ is such that $F(u_0, 0) = 0$ and $DF(u_0, 0) = Df(u_0)$. Because of the hyperbolicity, the linear transformation $Df(u_0)$ is invertible. By an application of the implicit function theorem, there is a smooth function $\beta\colon J \to \mathbb{R}^n$, where J is an open interval in $\mathbb{R}$ containing the origin, such that $F(\beta(\epsilon),\epsilon)) = 0$ for all ϵ in J. In other words, every small perturbation of (the vector field) f has a rest point near u_0. Moreover, for sufficiently small ϵ (and in view of the continuity of eigenvalues with respect to the components of the corresponding matrix), the perturbed rest point $\beta(\epsilon)$ is hyperbolic with the same number of eigenvalues (counting multiplicities) with positive and negative real parts as the eigenvalues of $Df(u_0)$. But, of course, the perturbed eigenvalues are in general not the same as the eigenvalues of $Df(u_0)$. The hyperbolic rest point u_0 is structurally stable in the sense that the local phase portraits at the corresponding rest points of sufficiently small perturbations of f are qualitatively the same as the phase portrait of the differential Eq. (23). This fact can be proved using the Grobman–Hartman theorem together with an argument to show that the phase portraits of hyperbolic linear first-order systems are topologically equivalent if their system matrices have the same numbers of eigenvalues (counting multiplicities) with positive and negative real parts. Suppose there were a diffeomorphism h taking orbits to orbits. It would correspond to a change of variables $v = h(u)$ (a smooth conjugacy) taking the differential Eq. (23) to the differential equation $\dot{v} = F(v)$, where the parameter ϵ is suppressed for simplicity. By differentiation of the equation $v = h(u)$ with respect to the independent variable t, it follows that $F(h(u)) = Dh(u)f(u)$. By linearization at the rest point u_0, we have

$$DF(h(u_0))Dh(u_0) = Dh(u_0)Df(u_0)\,.$$

In particular the system matrix $Df(h(u_0))$ for the linearized system at the perturbed rest point is similar, via the invertible matrix $Dh(u_0)$, to the system matrix $Df(u_0)$ for the linearized system at the unperturbed rest point. Thus, the two rest points would have to have exactly the same eigenvalues, contrary to the general situation for which the eigenvalues of the perturbed linearization are different from the eigenvalues of the unperturbed linearization.

The notion of hyperbolicity is an essential hypothesis in structural stability theory; indeed, hyperbolic rest points are structurally stable. Likewise hyperbolic periodic orbits, defined to be those periodic orbits such that there is an associated hyperbolic Poincaré map are structurally stable. This notion can be extended to general invariant sets. The essential requirement is that all of the solutions of the linearized differential equations along the solutions in the invariant set, which are not initially in the direction of the invariant set, either grow or decay exponentially with uniform exponential growth rates over the entire invariant set.

The global theory of structural stability requires a global hypothesis to ensure that the rest points and period orbits are connected with orbits in general position.

Definition 29 Let A be an invariant set (a union of orbits) of an autonomous first-order differential equation, vector field, or discrete dynamical system. The stable manifold of A, denoted $W^s(A)$ is the set of all solutions that converge to A as $t \to \infty$. The unstable manifold $W^u(A)$ is the set of all solutions that converge to A as $t \to -\infty$.

The stable manifold theorem asserts that the stable and unstable manifolds for rest points and periodic orbits of smooth vector fields are indeed (immersed) smooth manifolds (see, for example, [37,38]).

Two immersed manifolds are called transversal if they do not intersect; or, if the sum of their tangent spaces at each point of their intersection is the tangent space of the ambient manifold.

While periodicity is the most familiar form of recurrence for dynamical systems, a more subtle notion of recurrence is required for structural stability theorems.

Definition 30 The point u is called a nonwandering point for an autonomous differential equation if, for each neighborhood U of u and each time $t_0 > 0$, there is some time $t > t_0$ such that at least one solution starting in U at time zero returns to U at time t. Likewise for a discrete dynamical system defined by a diffeomorphism h, the point u is nonwandering if for each neighborhood U of u there is some integer $k > 0$ such that $h^k(U) \cap U$ is not empty. The set of all nonwandering points is called the nonwandering set.

For vector fields on compact orientable two-dimensional manifolds, the classical structural stability theorem is due to M. Peixoto [70] and is a generalization of earlier work for vector fields in the plane by L. Pointryagin and A. Andronov [2,3].

Theorem 31 *A class C^1 vector field on a compact orientable two-dimensional manifold M is structurally stable if and only if*

(1) all rest points are hyperbolic,
(2) all periodic solutions are hyperbolic,
(3) stable and unstable manifolds of all pairs of saddle points (rest points whose linearizations have one positive and one negative eigenvalue) are disjoint,
(4) the nonwandering set consists of only rest points and periodic solutions.

Moreover, the structurally stable vector fields form an open and dense set in $\mathfrak{X}(M)$.

The notion of transversality in Peixoto's theorem is embodied in the property (3). The generalization of this property is an essential ingredient of the theory.

Definition 32 A dynamical system satisfies the strong transversality property if for every pair of points $\{u, v\}$ in its nonwandering set, the corresponding invariant manifolds $W^s(u)$ and $W^u(y)$ intersect transversally (at all their points of intersection).

Definition 33 A dynamical system satisfies Axiom A if its nonwandering set is hyperbolic and the periodic orbits are dense in this set.

The most important result of the theory is the structural stability theorem:

Theorem 34 *A class C^1 diffeomorphism is structurally stable if and only if it satisfies Axiom A and the strong transversality property.*

Joel Robbin [77] proved that a C^2 diffeomorphism is C^1 structurally stable if it satisfies Axiom A and the strong transversality property. This result was improved by Clark Robinson [81] by relaxing the C^2 requirement to C^1; he also proved the analogous result for differential equations [79,80]. Ricardo Mañé [65] proved the converse.

A simple example of a structurally stable diffeomorphism is the north-pole south-pole map of the Riemann sphere; it is given by $z \mapsto 2z$ for the complex variable z. There are two hyperbolic rest points, an unstable fixed point at $z = 0$ and a stable fixed point at $z = \infty$. All other points are attracted to ∞ under forward iteration and to zero under backward (inverse) iteration.

A more exotic example is provided by the hyperbolic toral automorphisms. To define these maps, note that a 2×2-integer matrix with determinant one determines a diffeomorphism of the plane that preserves the integer lattice. Thus, it projects to a map of the torus formed by the quotient space $\mathbb{R}^2/\mathbb{Z}^2$. In case the matrix is hyperbolic, the corresponding diffeomorphism on the torus has a dense set of periodic orbits and thus its nonwandering set is the entire torus. It can be proved that the linear hyperbolic structure of the transformation of the plane (that is, one stable and one unstable eigenspace) projects to a uniform hyperbolic structure on the entire torus; and, moreover, the strong transversality condition is satisfied. Thus, small perturbations of such a map have the same properties. This example was generalized by Dmitri Anosov and others to include a class of diffeomorphisms and vector fields in the continuous case now called Anosov dynamical systems. The most important example of a continuous flow with this property is the geodesic flow in the unit tangent bundle of a compact Riemannian manifold with negative sectional curvatures (see [4]).

Attractors

A natural outgrowth of the theory of stability for rest points and periodic solutions is its generalization to invariant sets. The fundamental objects of study are the attractors. Unfortunately, the definition of this concept is not universally accepted; different authors use different definitions. The following definition is perhaps the most popular:

Definition 35 A closed invariant set A for a dynamical system is called an attracting set if it is contained in an open neighborhood U of the ambient space such that the intersection of the sequence of all forward motions of U under the action of the dynamical system is equal to A. An attracting set is called an attractor if, in addition, there is no closed invariant attracting subset of A.

The existence of attractors is part of the stability theory of dynamical systems; the structure of attractors is the subject of another rich theory. For most of the history of differential equations the only known attractors (the classical attractors) were rest points, periodic orbits, and tori, which are all manifolds. The existence of attractors with fractal dimensions, called strange attractors, was noticed and proved only recently [62,82,85,95].

For differential equations in $\mathbb{R}^n$, the existence of an attractor for autonomous systems is usually proved by demonstrating the existence of an $(n-1)$-dimensional sphere (or other separating hypersurface) such that the vector field corresponding to the differential equation points into the bounded region of space bounded by the sphere (more precisely, the inner product of the vector field and the inner normal is positive at all points on the sphere).

As an example, consider the Lorenz equations on $\mathbb{R}^3$

$$\begin{aligned} \dot{x} &= \sigma(y-x), \\ \dot{y} &= rx - y - xz, \\ \dot{z} &= xy - bz, \end{aligned}$$

where σ, r, and b are positive constants. The origin is a rest point for the Lorenz system and, if $r < 1$, then

$$L(x,y,z) := \frac{1}{\sigma}x^2 + y^2 + z^2$$

is a Lyapunov function. In this case, the origin is globally asymptotically stable. More generally (that is, with no restriction on the size of r), there is a constant $c > 0$ such that the vector field points into the bounded region of space bounded by the ellipsoid $rx^2 + \sigma y^2 + \sigma(z - 2r)^2 = c$. Thus, the Lorenz equations have an attractor. For some parameter values, for example $\sigma = 10$, $b = 8/3$, and $r = 28$, the attractor is a fractal, called the Lorenz attractor (see [29,62,93,95]).

For the special case of two-dimensional first-order autonomous systems, the existence and nature of attractors can be more precisely specified. The main result is the Poincaré-Bendixson theorem (see [15] and, for a proof, [1]):

Theorem 36 *An attractor of the class C^1 differential Eq. (23) on the plane that contains no rest points is a periodic orbit.*

Corollary 37 *If the differential Eq. (23) on the plane is analytic and has a positively invariant annulus containing no rest points, then the annulus contains an orbitally asymptotically stable periodic solution (a stable limit cycle).*

As an example, consider the model of a damped pendulum with torque given by the first-order planar system

$$\dot{\theta} = \nu \,, \quad \dot{\nu} = -\sin\theta + \mu - \lambda\nu \,, \tag{31}$$

where λ and μ are positive constants. By viewing the variable θ as an angular variable modulo 2π, the phase space is the cylinder rather than the plane. Nonetheless, the Poincaré-Bendixson theory is valid on the cylinder because a bounded region of the cylinder can be flattened by a change of variables to a region of the plane. If $|\mu| > 1$, then the system of differential Eq. (31) has a globally attracting periodic orbit. Three main observations can be used to construct a proof: there are no rest points, the quantity $-\sin\theta + \mu - \lambda\nu$ is negative for sufficiently large values of $\nu > 0$, and it is positive for negative values of ν with sufficiently large absolute values. These facts imply the existence of an invariant annulus containing no rest points. The remainder of the proof uses Theorem 19 and the analyticity of solutions (see p. 98 in [19]).

Generalizations and Future Directions

The concepts of stability theory have been successfully generalized to infinite-dimensional dynamical systems (which arise from the analysis of partial differential equations) defined on Banach spaces and to abstract dynamical systems on metric spaces [13,32,33,36,83,94]. Research in this direction has produced some results on stability and the existence of attractors.

No general theorem akin to Lyapunov's Theorem 4 is known for infinite-dimensional dynamical systems. A main difficulty is that, for an unbounded operator A, the non-zero elements in the spectrum of e^A may not all be given by the exponentials of elements in the spectrum of A (see, for a discussion, Chap. 2 in [20]). Thus, to determine the stability of a rest point by linearization, this spectral mapping property must be checked separately for most cases of interest.

Part of the motivation for the determination of attractors in infinite-dimensional dynamical systems is the possibility of reducing the long-term dynamical properties to the analysis of a finite-dimensional dynamical system on the attractor. The fundamental organizing concept in this direction is the "inertial manifold".

Definition 38 An inertial manifold is a finite-dimensional Lipschitz manifold that is invariant under forward motions of the dynamical system and attracts all solutions of the dynamical system at an exponential rate.

Inertial manifolds have been proved to exist for many dynamical systems defined by evolution-type partial differential equations, for example, reaction-diffusion equations with appropriate boundary conditions. Much current research is devoted to determining conditions that imply the existence of inertial manifolds (or their generalizations) for the equations of fluid dynamics.

The conceptual framework and basic abstract theory of stability is mature and well-understood for finite-dimensional dynamical systems. This theory is currently being extended to an abstract theory for infinite-dimensional dynamical systems. The most important direction for further development is the application of this theory to specific differential equations (including partial differential equations) that arise as mathematical models in applied mathematics. Even the problem that originally motivated stability theory (that is, the stability of periodic and quasi-periodic motions of the N-body problem of classical mechanics) remains open to future research.

Bibliography

Primary Literature

1. Alligood KT, Sauer TD, Yorke JA (1997) Chaos: An introduction to dynamical systems. Springer, New York
2. Andronov AA, Vitt EA, Khaiken SE (1966) Theory of Oscillations. Pergamon Press, Oxford
3. Andronov AA, Leontovich EA, Gordon II, Maier AG (1973) Qualitative Theory of Second-Order Dynamic Systems. Wiley, New York
4. Anosov DV (1967) Geodesic Flows on Closed Riemannian Manifolds with Negative Curvature. Proc Steklov Inst Math 90:1–209
5. Arnold VI (1963) Proof of an Kolmogorov's theorem on the preservation of quasi-periodic motions under small perturbations of the Hamiltonian. Usp Mat Nauk SSSR 113(5):13–40
6. Arnold VI (1964) Instability of dynamical systems with many degrees of freedom. Soviet Math 5(3):581–585
7. Arnold VI (1973) Ordinary Differential Equations. MIT Press, Cambridge
8. Arnold VI (1978) Mathematical Methods of Celestial Mechanics. Springer, New York
9. Arnold VI (1982) Geometric Methods In: The Theory of Ordinary Differential Equations. Springer, New York
10. Arnold VI (ed) (1988) Dynamical Systems I. Springer, New York
11. Arnold VI (ed) (1988) Dynamical Systems III. Springer, New York
12. Arscott FM (1964) Periodic Differential Equations. MacMillan, New York
13. Babin AV, Vishik MI (1992) Attractors of Evolution Equations. North-Holland, Amsterdam

14. Bautin NN (1954) On the number of limit cycles which appear with the variation of coefficients from an equilibrium position of focus or center type. Am Math Soc Transl 100:1–19
15. Bendixson I (1901) Sur les coubes définiés par des équations différentielles. Acta Math 24:1–88
16. Brjuno AD (1971) Analytic form of differential equations. Trudy MMO 25:119–262
17. Chang K-C (2005) Methods in Nonlinear Analysis. Springer, New York
18. Chetayev NG (1961) The Stability of Motion. Pergamon Press, New York
19. Chicone C (2006) Ordinary Differential Equations with Applications, 2nd edn. Springer, New York
20. Chicone C, Latushkin Y (1999) Evolution Semigroups. In: Dynamical Systems and Differential Equations. Am Math Soc, Providence
21. Chicone C, Liu W (2004) Asymptotic phase revisited. J Diff Eq 204(1):227–246
22. Chicone C, Swanson R (2000) Linearization via the Lie derivative. Electron J Diff Eq Monograph 02:64
23. Dirichlet GL (1846) Über die Stabilität des Gleichgewichts. J Reine Angewandte Math 32:85–88
24. Dirichlet GL (1847) Note sur la stabilité de l'équilibre. J Math Pures Appl 12:474–478
25. Dulac H (1908) Déterminiation et intégratin d'une certaine classe d'equatins différentielles ayant pour point singulier un centre. Bull Soc Math Fr 32(2):230–252
26. Dumortier F (2006) Asymptotic phase and invariant foliations near periodic orbits. Proc Am Math Soc 134(10):2989–2996
27. Floquet G (1883) Sur les équations différentielles linéaires à coefficients péiodiques. Ann Sci École Norm Sup 12(2):47–88
28. Frommer M (1934) Über das Auftreten von Wirbeln und Strudeln (geschlossener und spiraliger Integralkurven) in der Umgebung rationaler Unbestimmtheitsstellen. Math Ann 109:395–424
29. Guckenheimer J, Holmes P (1986) Nonlinear Oscillations, Dynamical Systems, and Bifurcation of Vector Fields, 2nd edn. Springer, New York
30. Grobman D (1959) Homeomorphisms of systems of differential equations. Dokl Akad Nauk 128:880–881
31. Grobman D (1962) Topological classification of the neighborhood of a singular point in *n*-dimensional space. Mat Sb 98:77–94
32. Hale JK (1988) Asymptotic Behavior of Dissipative Systems. Am Math Soc, Providence
33. Hale JK, Magalhães LT, Oliva WM (2002) Dynamics in Infinite Dimensions, 2nd edn. Springer, New York
34. Hartman P (1960) A lemma in the theory of structural stability of differential equations. Proc Am Math Soc 11:610–620
35. Hartman P (1960) On local homeomorphisms of Euclidean space. Bol Soc Math Mex 5(2):220–241
36. Henry D (1981) Geometric Theory of Semilinear Paabolic Equations. Lecture Notes in Mathematics, vol 840. Springer, New York
37. Hirsch M, Pugh C (1970) Stable manifolds and hyperbolic sets In: Global Analysis XIV. Am Math Soc, Providence, pp 133–164
38. Hirsch MC, Pugh C, Shub M (1977) Invariant Manifolds. Lecture Notes in Mathematics, vol 583. Springer, New York
39. Hofer H, Zehnder E (1994) Symplectic invariants and Hamiltonian dynamics. Birkhäuser, Basel
40. Hurwitz A (1895) Über die Bedingungen, unter welchen eine Gleichung nur Wurzeln mit negativen reellen Theilen besitzt. [On the conditions under which an equation has only roots with negative real parts.] Math Ann 46:273–284; (1964) In: Ballman RT et al (ed) Selected Papers on Math Trends in Control Theory. Dover, New York (reprinted)
41. Ilyashenko JS (1972) Algebraic unsolvability and almost algebraic solvability of the problem for the center-focus. Funkcional Anal I Priložen 6(3):30–37
42. Jost R (1968) Winkel- und Wirkungsvariable für allgemeine mechanische Systeme. Helvetica Phys Acta 41:965–968
43. Kapteyn W (1911) On the centra of the integral curves which satisfy differential equations of the first order and the first degree. Proc Kon Akak Wet Amsterdam 13:1241–1252
44. Kapteyn W (1912) New researches upon the centra of the integrals which satisfy differential equations of the first order and the first degree. Proc Kon Akak Wet Amsterdam 14: 1185
45. Kapteyn W (1912) New researches upon the centra of the integrals which satisfy differential equations of the first order and the first degree. Proc Kon Akak Wet Amsterdam 15:46–52
46. Kolmogorov AN (1954) La théorie générale des systèmes dynamiques et la méchanique. Amsterdam Congress I, pp 315–333
47. Kolmogorov AN (1954) On the conservation of conditionally periodic motions under small perturbations of the Hamiltonian. Dokl Akad Nauk SSSR 98:527–530
48. Lagrange JL (1776) Nouv Mem Acad Roy Sci. Belles-Lettres de Berlin, p 199
49. Lagrange JL (1788) Mechanique Analitique. Desaint, Paris, pp 66–73, pp 389–427; (1997) Boston Studies in the Philosophy of Science 191. Kluwer, Dordrecht (reprinted)
50. Lagrange JL (1869) Oeuvres de Lagrange. Gauthier, Paris 4:255
51. LaSalle JP (1976) The Stability of Dynamical Systems. SIAM, Philadelphia
52. Laplace PS (1784) Mem Acad Roy Sci. Paris, p 1
53. Laplace PS (1895) Oeuvres Complètes de Laplace. Gauthier, Paris 11:47
54. Leontovich AM (1962) On the stability of the Lagrange periodic solutions for the reduced problem of three bodies. Sov Math Dokl 3(2):425–430
55. Liapounoff AM (1947) Problème général de la stabilité du movement. Princeton University Press, Princeton
56. Liouville J (1842) Extrait d'une lettre à M Arago sur le Mémoire de M Maurice relatif à l'invariabilité des grands axes des orbites planétaires. Compt Rend Séances l'Acad Sci 15:425–426
57. Liouville J (1842) Extrait d'un Mémoire sur un cas particulier du problème des trois corps. J Math Sér I 7:110–113
58. Liouville J (1843) Sur la loi de la pesanteur à la surface ellipsoïdale d'équilibre d'une masse liquide homogène douée d'un mouvement de rotation. J Math Sér I 8:360
59. Liouville J (1855) Formules générales relatives à la question de la stabilité de l'équilibre d'une masse liquide homogène douée d'un mouvement de rotation autour d'un axe. Sér I, J Math Pures Appl 20:164–184
60. Lochak P (1999) Arnold diffusion; a compendium of remarks and questions. Hamiltonian systems with three or more degrees of freedom (S'Agaró 1995). In: NATO Adv Sci Inst Ser C Math Phys Sci, 533. Kluwer, Dordrecht, pp 168–183

61. Lochak P, Meunier C (1988) Multiphase averaging for classical systems. In: Trans H.S. Dumas. Springer, New York
62. Lorenz EN (1963) Deterministic nonperiodic flow. J Atm Sci 20:130–141
63. Lyapunov AM (1892) The General Problem of the Stability of Motion. Russ Math Soc, Kharkov (russian); (1907) Ann Facult Sci l'Univ Toulouse 9:203–474 (french trans: Davaux É); (1949) Princeton University Press, Princeton (reprinted); (1992) Taylor & Francis, London (english trans: Fuller AT)
64. Magnus W, Winkler S (1979) Hill's Equation. Dover, New York
65. Mañé R (1988) A proof of the C^1 stability conjecture. Inst Hautes Études Sci Publ Math 66:161–210
66. Marcus L, Yamabe H (1960) Global stability criteria for differential equations. Osaka Math J 12:305–317
67. Melnikov VK (1963) On the stability of the center for time periodic perturbations. Trans Mosc Math Soc 12:1–57
68. Moser J (1962) On invariant curves of area-preserving mappings of an annulus. Nachr Akad Wiss Göttingen, Math Phys K1:1–10
69. Moser J (1978/79) Is the solar system stable? Math Intell 1: 65–71
70. Peixoto MM (1962) Structural stability on two-dimensional manifolds. Topology 1:101–120
71. Poincaré H (1881) Mémoire sur les courbes définies par une équation différentielle. J Math Pures Appl 7(3):375–422; (1882) 8:251–296; (1885) 1(4):167–244; (1886) 2:151–217; all reprinted (1928) Tome I. Gauthier–Villar, Oeuvre, Paris
72. Poincaré H (1890) Sur le problème des trois corps et les équations del la dynamique. Acta Math 13:83–97
73. Poincaré H (1892) Les méthodes nouvelles de la mécanique céleste. Tome III. Invariants intégraux Solutions périodiques du deuxième genre, Solutions doublement asymptotiques. Dover, New York (1957)
74. Poincaré H (1951) Oeuvres, vol 1. Gauthier-Villans, Paris
75. Poincaré H (1957) Les méthodes nouvelles de la mécanique céleste. Tome I. Solutions périodiques. Non-existence des intègrales uniformes, Solutions asymptotiques. Dover, New York
76. Poincaré H (1957) Les méthodes nouvelles de la mécanique céleste. Tome II. Méthodes de MM. Newcomb, Gyldén, Lindstedt et Bohlin. Dover, New York
77. Robbin JW (1971) A structural stability theorem. Ann Math 94(2):447–493
78. Robbin JW (1995) Matrix algebra. A.K. Peters, Wellesley
79. Robinson C (1974) Structural stability of vector fields. Ann Math 99(2):154–175
80. Robinson C (1975) Errata to: "Structural stability of vector fields". (1974) Ann Math 99(2):154–175, Ann Math 101(2):368
81. Robinson C (1976) Structural Stability for C^1 diffeomorphisms. J Diff Eq 22(1):28–73
82. Robinson C (1989) Homoclinic bifurcation to a transitive attractor of Lorenz type. Nonlinearity 2(4):495–518
83. Robinson J (2001) Infinite Dimensional Dynamical Systems. Cambridge Univ Press, Cambridge
84. Routh EJ (1877) A Treatise on the Stability of a Given State of Motion. Macmillan, London; Fuller AT (ed) (1975) Stability of Motion. Taylor, London (reprinted)
85. Rychlik MR (1990) Lorenz attractors through Šil'nikov-type bifurcation I. Ergodic Theory Dynam Syst 10(4):793–821
86. Sanders JA, Verhulst F, Murdock J (2007) Averaging Methods. In: Nonlinear Dynamical Systems, 2nd edn. Appl Math Sci 59. Springer, New York
87. Siegel CL (1942) Iterations of anayltic functions. Ann Math 43:607–612
88. Siegel CL (1952) Über die Normalform analytischer Differentialgleichungen in der Nähe einer Gleichgewichtslösung. Nachr Akak Wiss Gottingen, Math Phys K1:21–30
89. Siegel CL, Moser J (1971) Lectures on Celestial Mechanics. Springer, New York
90. Sternberg S (1958) On the structure of loacal homeomorphisms of Euclidean *n*-space. Am J Math 80(3):623–631
91. Sternberg S (1959) On the structure of loacal homeomorphisms of Euclidean *n*-space. Am J Math 81(3):578–604
92. Sternberg S (1969) Celestial Mechanics. Benjamin, New York
93. Strogatz SH (1994) Nonlinear Dynamics and Chaos. Perseus Books, Cambridge
94. Temam R (1997) Infinite-Dimensional Dynamical Systems. In: Mechanics and Physics. Springer, New York
95. Viana M (2000) What's new on Lorenz strange attractors? Math Intell 22(3):6–19
96. Wiggins S (2003) Introduction to Applied Nonlinear Dynamical Systems and Chaos, 2nd edn. Springer, New York
97. Willems JL (1970) Stability Theory of Dynamical Systems. Wiley, New York
98. Yakubovich VA, Starzhinskii VM, Louvish D (trans) (1975) Linear Differential Equations with Periodic Coefficients. Wiley, New York
99. Żołądek H (1994) Quadratic systems with center and their perturbations. J Diff Eq 109(2):223–273
100. Żołądek H (1997) The problem of the center for resonant singular points of polynomial vector fields. J Diff Eq 135:94–118

Books and Reviews

Arnold VI, Avez A (1968) Ergodic Problems of Classical Mechanics. Benjamin, New York

Coddington EA, Levinson N (1955) Theory of ordinary differential equations. McGraw-Hill, New York

Farkas M (1994) Periodic Motions. Springer, New York

Glendinning PA (1994) Stability, Instability and Chaos. Cambridge University Press, Cambridge

Grimshaw R (1990) Nonlinear Ordinary Differential Equations. Blackwell, Oxford

Hale JK (1980) Ordinary Differential Equations, 2nd edn. RE Krieger, Malabar

Hahn W, Baartz AP (trans) (1967) Stability of Motion. Springer, New York

Hartman P (1964) Ordinary Differential Equations. Wiley, New York

Hirsch M, Smale S (1974) Differential Equations, Dynamical Systems, and Linear Algebra. Acad Press, New York

Katok AB, Hasselblatt B (1995) Introduction to the Modern Theory of Dynamical Systems. Cambridge Univ Pres, Cambridge

Krasovskiĭ NN (1963) Stability of Motion. Stanford Univ Press, Stanford

LaSalle J, Lefschetz S (1961) Stability by Lyapunov's Direct Method with Applications. Acad Press, New York

Lefschetz S (1977) Differential Equations: Geometric Theory, 2nd edn. Dover, New York

Miller RK, Michel AN (1982) Ordinary Differential Equations. Acad Press, New York

Moser J (1973) Stable and Random Motions. In: Dynamical Systems Ann Math Studies 77. Princeton Univ Press, Princeton

Nemytskiĭ VV, Stepanov VV (1960) Qualitative Theory of Differential Equations. Princeton Univ Press, Princeton

Perko L (1996) Differential Equations and Dynamical Systems, 2nd edn. Springer, New York

Robbin JW (1970) On structural stability. Bull Am Math Soc 76:723–726

Robbin JW (1972) Topological conjugacy and structural stability for discrete dynamical systems. Bull Am Math Soc 78(6):923–952

Robinson C (1995) Dynamical Systems: Stability, Symbolic Dynamics, and Chaos. CRC Press, Boca Raton

Smale S (1967) Differentiable dynamical systems. Bull Am Math Soc 73:747–817

Smale S (1980) The Mathematics of Time: Essays on Dynamical Systems, Economic Processes and Related Topics. Springer, New York

Sotomayor J (1979) Liçóes de equaçóes diferenciais ordinárias. IMPA, Rio de Janerio

Verhulst F (1989) Nonlinear Differential Equations and Dynamical Systems. Springer, New York

Static Games

Oscar Volij
Ben-Gurion University, Beer-Sheva, Israel

Article Outline

Glossary

Player A participant in a game.

Action set The set of actions that a player may choose.

Action profile A list of actions, one for each player.

Payoff The utility a player obtains from a given action profile.

Definition of the Subject

Game theory concerns the interaction of decision makers. This interaction is modeled by means of games. There are various approaches to constructing games. One approach is to focus on the possible outcomes of the decision-makers' interaction by abstracting from the actions or decisions that may lead to these outcomes. The main tool used to implement this approach is the *cooperative game*. Another approach is to focus on the actions that the decision-makers can take, the main tool being the *non-cooperative game*. Within this approach, strategic interactions are modeled in two ways. One is by means of *dynamic*, or extensive form games, and the other is by means of *static*, or strategic games. Dynamic games stress the sequentiality of the various decisions that agents can make. An essential component of a dynamic game is the description of who moves first, who moves second, etc. Static games, on the other hand, abstract from the sequentiality of the possible moves, and model interactions as simultaneous decisions, where the decisions may well be complicated plans of actions that dictate different moves for different situations that may arise. All extensive form games can be modeled as static games, and all strategic form games can be modeled as extensive form games. But some situations may be more conveniently modeled as one or the other kind of game.

This chapter reviews the main ideas and results related to static games, as well as some interesting relationships that connect equilibrium concepts with the idea of rationality. The objective is to introduce the reader to the area of static games and to stimulate his interest for further knowledge of game theory in general. For a comprehensive exposition of some results not covered in this chapter, the reader is referred to the many excellent textbooks available on game theory. Binmore [6], Fudenberg and Tirole [9], Osborne [20], Osborne and Rubinstein [21] constitute only a partial list.

Although the definition of a static game is a very simple one, static games are a very flexible model which allows us to analyze many different situations. In particular, one can use them to analyze strategic interactions that involve either common interests or diametrically opposed interests. Similarly, one can also use static games to model situations where players have either symmetric or asymmetric information. The range of applications of static games is very wide and covers many disciplines, such as economics, political science, biology, philosophy, and computer science among others.

Introduction

In this section we introduce some examples that will be used later to motivate different concepts. We also introduce the definition of a static game.

The prisoner's dilemma involves a donor who is interested in donating some amount of money to two universities. The donor decides that the amount each university will receive depends on the content of the messages the presidents of the respective universities will send to him. Each university will send simultaneously one of two messages. One possible message is "Give him 2" and the other is "Give me 1". The donor will do exactly as told. For instance, if University I sends the message "Give me 1" and University II sends "Give him 2", the donor will donate \$3 to University I and \$0 to University II. This game can be described by means of the following matrix, where the entries represent the payoffs for University I and University II, respectively, that result from the corresponding action choices.

		University II	
		Give him 2	*Give me 1*
University I	*Give him 2*	2, 2	0, 3
	Give me 1	3, 0	1, 1

The battle of the sexes consists of two friends, She and He, who want to go out together, but have no means of communication. They have to decide, each one separately but both simultaneously, whether to go to a boxing match or to a ballet show. For both of them, the worst possible outcome would be to choose different events and not meet. But if they meet, he would rather meet her at the boxing match, while she would rather meet him at the ballet. The battle of the sexes can be described by the following matrix.

		She	
		Box	Ballet
He	Box	2, 1	0, 0
	Ballet	0, 0	1, 2

Again, the entries of this matrix represent the payoffs that he and she get, as a result of their corresponding choices.

Chicken models two drivers who approach each other on a narrow street. If none of them slows down they'll have an accident and their corresponding payoffs will be 0. But if at least one of them slows down, the accident is prevented. The problem is that both of them would like the other to slow down. If only one driver slows down, this driver gets a payoff of 2 and the other driver gets a payoff of 7. If both drivers slow down, then both drivers get a payoff of 6. This situation can be described by the following matrix.

		Driver 2	
		Slow Down	Speed up
Driver 1	Slow Down	6, 6	2, 7
	Speed up	7, 2	0, 0

Matching Pennies involves two friends, each of whom places a coin on a table. If both coins are placed heads up or tails up, then friend 1 gets one dollar from friend 2. If one coin is placed heads up and the other tails up, then friend 1 pays one dollar to friend 2. Matching pennies can be described by the following matrix, where the entries are the amounts of money that the friends get from each other.

		Friend 2	
		Heads	Tails
Friend 1	Heads	$1, -1$	$-1, 1$
	Tails	$-1, 1$	$1, -1$

The above examples of strategic interactions can be modeled as static games. A static game is a formalization of a strategic situation according to which players choose their actions separately and simultaneously, and as a result obtain certain payoffs. The interaction that a static game models need not require that players take their actions simultaneously. But the interaction is modeled by defining actions in such a way that lets us think of the players as acting simultaneously.

All of the above examples involve a set of players, and for each player there is a set of available actions and a function that associates a payoff level to each of the profiles of actions that may result from the players' choices. These are the three essential components of a static game, as formalized in the following definition.

Definition 1 A static game is a triple $\langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$ where N is a finite set of players, and for each player $i \in N$, A_i is i's set of actions, and $u_i : \times_{k \in N} A_k \rightarrow \mathbb{R}$ is player i's utility function.

In the prisoner's dilemma the set of players is $N = \{\text{University I, University II}\}$; the sets of actions are $A_{\text{I}} = A_{\text{II}} = \{\textit{Give me 1}, \textit{Give him 2}\}$; the utility function of University I is $u_{\text{I}}(\textit{Give me 1}, \textit{Give me 1}) = 1$, $u_{\text{I}}(\textit{Give me 1}, \textit{Give him 2}) = 3$, $u_{\text{I}}(\textit{Give him 2}, \textit{Give me 1}) = 0$, $u_{\text{I}}(\textit{Give him 2}, \textit{Give him 2}) = 2$; and the utility function of University II is $u_{\text{II}}(\textit{Give me 1}, \textit{Give me 1}) = 1$, $u_{\text{II}}(\textit{Give me 1}, \textit{Give him 2}) = 0$, $u_{\text{II}}(\textit{Give him 2}, \textit{Give me 1}) = 3$, $u_{\text{I}}(\textit{Give him 1}, \textit{Give him 1}) = 1$.

In this chapter we sometimes refer to static games simply as games. For any game $\langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$, the set of action profiles $\times_{k \in N} A_k$ is denoted by A, and a typical action profile is denoted by $a = (a_i)_{i \in N} \in A$. If A is a finite set, then we say that the game is finite. Player i's utility function represents his preferences over the set of action profiles. For instance, for any two action profiles a and a' in A, $u_i(a) \geq u_i\left(a'\right)$ means that player i prefers action profile a to action profile a'. Clearly, although player i has preferences over action profiles, he can only affect his own component, a_i, of the profile.

Nash Equilibrium

One objective of game theory is to select, for each game, a set of action profiles that are interesting in some way. These action profiles may be interpreted as predictions of the theory, or prescriptions for the players to follow, or simply as equilibrium outcomes in the sense that if they occur, the players do not wish that they had acted differently. These action profiles are formally given by solution concepts, which are functions that associate each strategic game with the selected set of action profiles. The central solution concept in game theory is known as *Nash equilibrium*. The hypothesis behind this solution concept is that each player chooses his actions so as to maximize his utility, given the profile of actions chosen by the other players. To give a formal definition of the Nash equilibrium concept, we first introduce some useful notation. For each player $i \in N$, let $A_{-i} = \times_{k \in N \setminus \{i\}} A_k$ be the set of the other players' profiles of actions. Then we can write $A = A_i \times A_{-i}$, and each action profile can be written as $a = (a_i, a_{-i}) \in A_i \times A_{-i}$, thereby distinguishing player i's action from the other players' profile of actions.

Definition 2 The action profile $a^* = (a_i^*)_{i \in N} \in A$ in a game $\langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$ is a *Nash equilibrium* if for each player, $i \in N$, and every action $a_i \in A_i$ of player i, a^* is at least as good for player i as the action profile (a_i, a_{-i}^*). That is, if

$$u_i(a^*) \geq u_i(a_i, a_{-i}^*) \quad \text{for all } a_i \in A_i \quad \text{and for all } i \in N\,.$$

It is a *strict Nash equilibrium* if the above inequality is strict for all alternative actions $a_i \in A_i \setminus \{a_i^*\}$.

Analysis of Some Finite Games

Prisoner's Dilemma Recall that the prisoner's dilemma can be described by the following matrix.

		University II	
		Give him 2	*Give me 1*
University I	*Give him2*	2, 2	0, 3
	Give me1	3, 0	1, 1

The action profile (*Give me 1, Give me 1*) is a Nash equilibrium. Indeed,

$$u_{\mathrm{I}}(\textit{Give me 1, Give me 1}) = 1 \\ \geq u_{\mathrm{I}}(\textit{Give him } 2,\ \textit{Give me } 1) = 0$$

and

$$u_{\mathrm{II}}(\textit{Give me } 1,\ \textit{Give me } 1) = 1 \\ \geq u_{\mathrm{II}}(\textit{Give me } 1,\ \textit{Give him } 2) = 0\,.$$

On the other hand, the action profile (*Give him 2, Give him 2*) is not a Nash equilibrium, since University I prefers action "*Give me 1*" if University II chooses action "*Give him 2*":

$$2 = u_{\mathrm{I}}(\textit{Give him } 2,\ \textit{Give him } 2) \\ < u_{\mathrm{I}}(\textit{Give me } 1,\ \textit{Give him } 2) = 3\,.$$

Battle of the Sexes Recall that the battle of the sexes can be described by the following matrix.

		She	
		Box	Ballet
He	Box	2, 1	0, 0
	Ballet	0, 0	1, 2

One can check that (*Box, Box*) is a Nash equilibrium and (*Ballet, Ballet*) is a Nash equilibrium as well. It can also be checked that these are the only two action profiles that constitute a Nash equilibrium.

Matching Pennies The reader can check that Matching Pennies has no Nash equilibrium.

Before we analyze the next example, we introduce a technical tool that allows us to reformulate the definition of Nash equilibrium more conveniently. More importantly, this alternative definition is the key to the standard proof of the existence of Nash equilibrium.

Definition 3 Let $G = \langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$ be a strategic game and let $i \in N$ be a player. Consider a list of actions $a_{-i} = (a, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n) \in \times_{k \in N \setminus \{i\}} A_k$ of all the players other than i. The set of player i's best responses to a_{-i} is

$$\mathcal{B}_i(a_{-i}) = \{a_i \in A_i \colon u_i(a_i, a_{-i}) \geq u_i(b_i, a_{-i}) \\ \text{for all } b_i \in A_i\}\,.$$

The correspondence $\mathcal{B}_i \colon \times_{j \neq i} A_j \twoheadrightarrow A_i$ that assigns to each $(n-1)$-tuple of actions in A_{-i} the set of best responses to it is called the *best response correspondence* of player i.

The definition of a Nash equilibrium may be stated in terms of the players' best response correspondences, as stated in the following proposition.

Proposition 1 *The action profile $a^* \in A$ is a Nash equilibrium if and only if every player's action is a best response to the other players' actions. That is, if*

$$a_i^* \in \mathcal{B}_i(a_{-i}^*) \quad \textit{for all } i \in N\,.$$

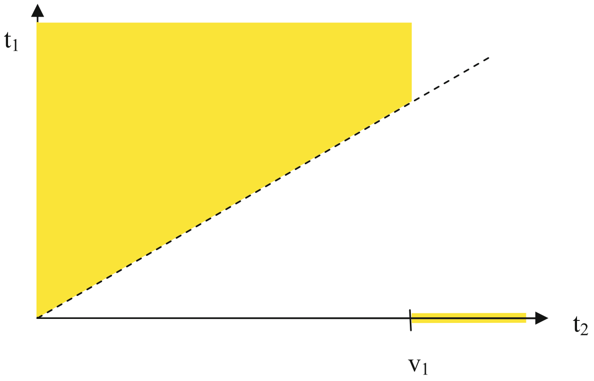
Until now, all the examples involved games where the action sets contained two actions. The next example is a game where the players' action sets are infinite. We will use the player's best response correspondences to find all its Nash equilibria.

The War of Attrition Two animals, 1 and 2, are fighting over a prey. Each animal chooses a time at which it intends to give up. Once one animal has given up, the other obtains the prey; if both animals give up at the same time then they split the prey equally. For each $i = 1, 2$, animal i's willingness to fight for the prey is given by $v_i > 0$. The value v_i is the maximum amount of time that animal i is willing to spend to obtain the prey. Since fighting is costly, each animal prefers as short a fight as possible. If animal i obtains the prey after a fight of length t, his utility will be $v_i - t$. We can model the situation as the game $G = \langle\{1,2\},(A_1,A_2),(u_1,u_2)\rangle$ where

- $A_1 = [0,\infty] = A_2$ (an element $t \in A_i$ represents a time at which player i plans to give up)
- $$u_1(t_1,t_2) = \begin{cases} -t_1 & \text{if } t_1 < t_2 \\ \frac{1}{2}v_1 - t_2 & \text{if } t_1 = t_2 \\ v_1 - t_2 & \text{if } t_1 > t_2 \end{cases}$$
- $$u_2(t_1,t_2) = \begin{cases} -t_2 & \text{if } t_2 < t_1 \\ \frac{1}{2}v_2 - t_1 & \text{if } t_1 = t_2 \\ v_2 - t_1 & \text{if } t_2 > t_1\,. \end{cases}$$

We are interested in the best response correspondences. First, we calculate player 1's best response correspondence, $\mathcal{B}_1(t_2)$. There are three cases to consider.

Case 1: $t_2 < v_1$ In this case, $v_1 - t_2 > \frac{1}{2}v_1 - t_2$ and $v_1 - t_2 > -t_1$. Consequently, given that player 2's action is t_2, player 1's utility function has a maximum value of $v_1 - t_2$, which is attained at any $t_1 > t_2$. Therefore, $\mathcal{B}_1(t_2) = (t_2,\infty)$.



Static Games, Figure 1
Player 1's best response correspondence

Case 2: $t_2 = v_1$ In this case, $0 = v_1 - t_2 > \frac{1}{2}v_1 - t_2$. Therefore, player's 1 utility function $u_1(\cdot,t_2)$ has a maximum value of 0, which is attained at $t_1 = 0$ and at $t_1 > t_2$. Therefore, $\mathcal{B}_1(t_2) = \{0\} \cup (t_2,\infty)$.

Case 3: $t_2 > v_1$ In this case $\frac{1}{2}v_1 - t_2 < v_1 - t_2 < 0$. As a result, player 1's utility function $u_1(\cdot,t_2)$ has a maximum value of 0, which is attained at $t_1 = 0$. Therefore, $\mathcal{B}_1(t_2) = \{0\}$.

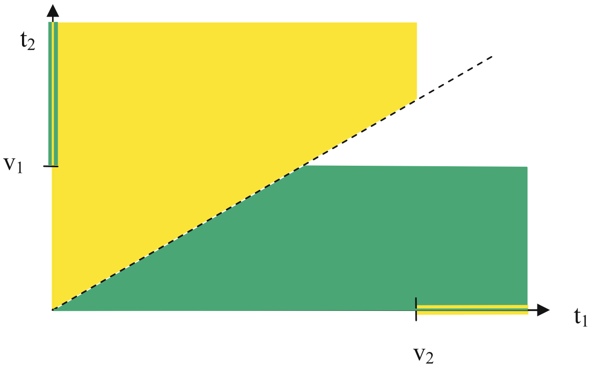
Summarizing, player 1's best response correspondence is:

$$\mathcal{B}_1(t_2) = \begin{cases} (t_2,\infty) & \text{if } t_2 < v_1 \\ \{0\} \cup (t_2,\infty) & \text{if } t_2 = v_1 \\ \{0\} & \text{if } t_2 > v_1 \end{cases}$$

which is depicted in Fig. 1.
Similarly, player 2's best response correspondence is:

$$\mathcal{B}_2(t_1) = \begin{cases} (t_1,\infty) & \text{if } t_1 < v_2 \\ \{0\} \cup (t_1,\infty) & \text{if } t_1 = v_2 \\ \{0\} & \text{if } t_1 > v_2\,. \end{cases}$$

Combining the two best response correspondences we get that (t_1^*, t_2^*) is a Nash equilibrium if and only if either $t_1^* = 0$ and $t_2^* \geq v_1$ or $t_2^* = 0$ and $t_1^* \geq v_2$. Figure 2 depicts the set of all the Nash equilibria as the intersection of the two best response correspondences. Two things are worth noting. First, it is not necessarily the case that the player who values the prey most wins the war. That is, there are Nash equilibria of the war of attrition where the player with the highest willingness to fight for the prey gives in first, and as a result the object goes to the other player. Second, in none of the Nash equilibria is there a physical fight. All Nash



Static Games, Figure 2
The equilibria

equilibria involve one player giving in immediately to the other. This second feature seems rather unrealistic, since fights in "war of attrition"-like situations are commonly observed. If one wants to obtain a fight of positive length in the war of attrition one needs to either drop the Nash equilibrium concept and adopt an alternative one, or model the war of attrition differently. We will adopt this second course of action later.

Existence

As the matching pennies example shows, not all games have a Nash equilibrium. The following theorem, which dates back to Nash [18] and Glicksberg [11], states sufficient conditions on a game for it to have a Nash equilibrium. An earlier version of this theorem for the smaller but prominent class of zero-sum games can be found in von Neumann [23] (translated in von Neumann [24]). The standard proofs use Kakutani's fixed point theorem. We present here an alternative proof, due to Geanakoplos [10], which uses Brouwer's fixed point theorem instead.

Theorem 1 *The game $\langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$ has a Nash equilibrium if for all $i \in N$*

- *the set A_i of actions of player i is a nonempty compact convex subset of an Euclidean space,*
- *the utility function u_i is continuous,*
- *the utility function u_i is concave in A_i.*

Proof (Geanakoplos) Define the correspondence $\varphi_i \colon A \twoheadrightarrow A_i$ by

$$\varphi_i(\bar{a}) = \arg\max_{a_i \in A_i} \{U_i(a_i, \bar{a}_{-i}) - \|a_i - \bar{a}_i\|^2\} ,$$

where, $\|\cdot\|$ denotes a norm in the relevant Euclidean space. Note first that φ_i is a nonempty valued correspondence because the maximand is a continuous function and A_i is compact. Second, note that the function $\|a_i - \bar{a}_i\|$ is convex:

$$\begin{aligned} &\|(\lambda a_i + (1-\lambda)b_i) - \bar{a}_i\| \\ &\quad = \|(\lambda a_i - \lambda \bar{a}_i) + ((1-\lambda)b_i - (1-\lambda)\bar{a}_i)\| \\ &\quad \leq \|(\lambda a_i - \lambda \bar{a}_i)\| + \|((1-\lambda)b_i - (1-\lambda)\bar{a}_i)\| \\ &\quad \leq |\lambda| \|a_i - \bar{a}_i\| + |1-\lambda| \|b_i - \bar{a}_i\| . \end{aligned}$$

Since the quadratic function is strictly convex, then the maximand is a strictly concave function. Therefore, the correspondence φ_i is in fact a function. Furthermore, since the maximand is continuous in the parameter $\bar{a}$, φ_i is also continuous. To see this, let $\bar{a}_n \to \bar{a}$ be a convergent sequence of action profiles and let $a_{i\,n} = \varphi_i(\bar{a}_n)$. This means that $U(a_{i\,n}, (\bar{a}_n)_{-i}) \geq U(b_i, (\bar{a}_n)_{-i})$ for all $b_i \in A_i$. Since A_i is a compact set, $a_{i\,n}$ has a convergent subsequence. Denoting by a_i the limit of this subsequence and applying limits to the above inequality, we obtain that

$$U(a_i, \bar{a}_{-i}) \geq U(b_i, \bar{a}_{-a}) \quad \text{for all } b_i \in A_i ,$$

namely $a_i = \varphi_i(\bar{a})$. Since this is true for every convergent subsequence of $a_{i\,n}$, we have that $\varphi_i(\bar{a}_n) = a_{i\,n} \to a_i = \varphi_i(\bar{a})$, which means that φ is continuous.

Now define $\varphi \colon A \to A$ by $\varphi = (\varphi_1, \ldots, \varphi_N)$. Clearly, φ is a continuous function mapping a compact set to itself. Therefore, by Brouwer's fixed point theorem, it has a fixed point: $\varphi(\bar{a}) = \bar{a}$. We now show that $\bar{a}$ is a Nash equilibrium of the game. Assume not. Then, there is some $i \in N$ with $a_i \in A_i$ such that $U_i(a_i, \bar{a}_{-i}) - U_i(\bar{a}) = E > 0$. Then, by concavity of U_i, for all $0 < \epsilon < 1$,

$$\begin{aligned} &U_i(\epsilon a_i + (1-\epsilon)\bar{a}_i, \bar{a}_{-i}) - U_i(\bar{a}) \\ &\quad \geq \epsilon U_i(a_i, \bar{a}_{-i}) + (1-\epsilon)U_i(\bar{a}) - U_i(\bar{a}) \\ &\quad \geq \epsilon E > 0 , \end{aligned}$$

while $\|\epsilon a_i + (1-\epsilon)\bar{a}_i - \bar{a}_i\|^2 = \epsilon^2 \|a_i - \bar{a}_i\|^2 < \epsilon E$, for small enough ϵ. Therefore, for such small ϵ, the action $\epsilon a_i + (1-\epsilon)\bar{a}_i$ satisfies

$$U_i(\epsilon a_i + (1-\epsilon)\bar{a}_i, \bar{a}_{-i}) - \|\epsilon a_i + (1-\epsilon)\bar{a}_i - \bar{a}_i\|^2 > U_i(\bar{a})$$

which contradicts the fact that $\varphi_i(\bar{a}) = \bar{a}_i$. □

Mixed Strategies

So far, we have formally defined a game, and have introduced the solution concept of Nash equilibrium which is arguably the central solution concept of game theory. However, there seem to be two problems with this concept. One is that although Nash equilibria exist in a wide class of games, there are many simple games that do not have a Nash equilibrium. The most troubling example is Matching Pennies. If game theory cannot provide a prediction for this simple game then one must wonder if there is any value to the theory. The second problem is that the concept of Nash equilibrium predicts a very unrealistic outcome in the war of attrition. One would expect that game theory would not only provide nonempty predictions, but also ones that look reasonable and help explain what we see around us.

One way to approach these problems is not to abandon the theory or the concept of Nash equilibrium altogether, but to modify the way we model the problematic situations. The idea behind mixed strategies is to first modify the game by extending the set of actions available to the players, and then to apply the concept of Nash equilibrium to this extended game. In this way one may obtain additional Nash equilibria, some of which may provide reasonable predictions to the game.

Let $G = \langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$ be a finite game. For any A_i, a probability distribution on A_i is a function

$$x_i \colon A_i \to \mathbb{R}_+$$

such that

$$\sum_{a_i \in A_i} x_i(a_i) = 1 \,.$$

The set of all probability distributions on A_i is denoted by $\Delta(A_i)$. A *mixed strategy on* A_i is a random choice over elements of A_i, namely an element of $\Delta(A_i)$. If x_i is a mixed strategy on A_i, $x_i(a_i)$ denotes the probability that action $a_i \in A_i$ is selected when x_i is adopted. Since elements of $\Delta(A_i)$ can have an alternative interpretation, such as beliefs about the choice of player i, we denote the set of mixed strategies by X_i to distinguish it from the more abstract set of probability distributions on A_i. Also, we denote the set of mixed strategy profiles as $X = \times_{i \in N} X_i$. Denoting for each player $i \in N$, $X_{-i} = \times_{k \in N \setminus \{i\}} X_k$, a typical mixed strategy profile can be written as $(x_k)_{k \in N} = (x_i, x_{-i}) \in X_i \times X_{-i}$. The *mixed extension* of the strategic game G is the strategic game $\langle N, (X_i)_{i \in N}, (U_i)_{i \in N} \rangle$ where the set of actions of player i is the set of mixed strategies, X_i, and the payoff function $U_i \colon \times_{i \in N} X_i \to \mathbb{R}$ of player i is defined by

$$U_i((x_k)_{k \in N}) = \sum_{a = (a_k)_{k \in N} \in A} u_i(a) \Pi_{k \in N} x_k(a_k) \,.$$

Remark 1 Since each mixed strategy of player i, x_i, can be identified with a vector $x_i = (x_i(a_i))_{a_i \in A_i} \in \mathbb{R}^{|A_i|}$, the function U_i is multinomial in the coordinates of its variables, and, as a result, it is continuous as a function of the players' mixed strategies.

Definition 4 An *equilibrium in mixed strategies* of the game $\langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$ is a Nash equilibrium of the mixed extension of the game. In other words, it is a list of mixed strategies $\left(x_k^*\right)_{k \in N} \in X$ such that for all players $i \in N$ and for all his mixed strategies x_i,

$$U_i\left(\left(x_k^*\right)_{k \in N}\right) \geq U_i\left(\left(x_i, x_{-i}^*\right)\right) \,.$$

Alternatively, $\left(x_k^*\right)_{k \in N} \in X$ is a mixed strategy equilibrium if

$$x_i^* \in \mathcal{B}_i\left(x_{-i}^*\right) \quad \text{for all } i \in N \,.$$

Note that for every finite game $G = \langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$, its mixed extension is a strategic game that satisfies the conditions of Theorem 1. As a result, every finite game has a mixed strategy equilibrium.

Example 1 Consider again Matching Pennies. Its mixed extension is the game $\langle N, (X_i)_{i \in N}, (U_i)_{i \in N} \rangle$ where the set of players is $N = \{1, 2\}$, the sets of mixed strategies are $X_1 = \{(p_H, p_T) \geq (0, 0) \colon p_H + p_T = 1\}$, and $X_2 = \{(q_H, q_T) \geq (0, 0) \colon q_H + q_T = 1\}$, and the utility functions are given by $U_1((p_H, p_T), (q_H, q_T)) = p_H q_H + p_T q_T - p_H q_T - p_T q_H$ and $U_2((p_H, p_T), (q_H, q_T)) = p_H q_T + p_T q_H - p_H q_H - p_T q_T$. It can be checked that the only Nash equilibrium of this mixed extension is $((1/2, 1/2), (1/2, 1/2))$. Indeed, since $U_1((p_H, p_T), (1/2, 1/2))$ is identically 0, it attains its maximum at, among other strategies, $(1/2, 1/2)$. The same is true for $U_2((1/2, 1/2), (q_H, q_T))$. To see that there is no other equilibrium, note that for (q_H, q_T) with $q_H > q_T$, player 1's best response is $(1, 0)$. But player 2's best response to $(1, 0)$, is $(0, 1)$. Since $0 \leq 1$, (q_H, q_T) with $q_H > q_T$ cannot be part of an equilibrium. Similarly, for any (q_H, q_T) with $q_H < q_T$, player 1's best response is $(0, 1)$. But player 2's best response to $(0, 1)$ is $(1, 0)$. Since $1 \geq 0$, (q_H, q_T) with $q_H < q_T$ cannot be part of an equilibrium.

We next present a characterization of the mixed strategy equilibria of a game that will sometimes allow us to compute them in an easy way. Further, this characterization serves as the basis of an interesting interpretation of the mixed strategy equilibrium concept that we will discuss later. For this purpose, we identify the action $a_i \in A_i$ of player i with the mixed strategy of player i that assigns probability 1 to action a_i, and 0 to all other actions. Therefore, given a player i, one of his actions $a_i \in A_i$, and a profile $x = (x_k)_{k \in N}$ of the players' mixed strategies, (a_i, x_{-i}) denotes the mixed strategy profile obtained from x by replacing i's mixed strategy x_i by the mixed strategy of player i that assigns probability 1 to action a_i. With this notation we can state the following identity:

$$U_i((x_k)_{k \in N}) = \sum_{a_i \in A_i} x_i(a_i) U_i((a_i, x_{-i})) . \qquad (1)$$

Indeed,

$$\begin{aligned} U_i((x_k)_{k \in N}) &= \sum_{a=(a_k)_{k \in N} \in A} u_i(a) \Pi_{k \in N} x_k(a_k) \\ &= \sum_{a_i \in A_i} \sum_{a_{-i} \in A_{-i}} u_i(a) \Pi_{k \in N} x_k(a_k) \\ &= \sum_{a_i \in A_i} x_i(a_i) \sum_{a_{-i} \in A_{-i}} u_i(a) \Pi_{k \in N \setminus \{i\}} x_k(a_k) \\ &= \sum_{a_i \in A_i} x_i(a_i) U_i((a_i, x_{-i})) . \end{aligned}$$

Identity (1) is useful to prove the following characterization of the mixed strategy Nash equilibria.

Lemma 1 *The strategy profile $x^* = (x_k^*)_{k \in N}$ is an equilibrium of the mixed extension of $\langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$ if and only if for all players $i \in N$ and for all $a_i \in A_i$,*

$$\text{If } x_i^*(a_i) > 0 \quad \text{then } U_i((a_i, x_{-i}^*)) = U_i(x^*) \qquad (2)$$

$$\text{If } x_i^*(a_i) = 0 \quad \text{then } U_i((a_i, x_{-i}^*)) \leq U_i(x^*) . \qquad (3)$$

Proof Assume that $x^* = (x_k^*)_{k \in N}$ satisfies conditions (2) and (3). Let $i \in N$, and let x_i be a mixed strategy of player i. Then, by (1)

$$\begin{aligned} U_i(x_i, x_{-i}^*) &= \sum_{a_i \in A_i} x_i(a_i) U_i((a_i, x_{-i}^*)) \\ &\leq \sum_{a_i \in A_i} x_i(a_i) U_i(x^*) = U_i(x^*) \end{aligned}$$

and therefore x^* is an equilibrium.

Assume now that $x^* = (x_k^*)_{k \in N}$ is an equilibrium. Let $i \in N$. Then

$$U_i(x^*) \geq U_i((a_i, x_{-i}^*)) \quad \forall a_i \in A_i \qquad (4)$$

and, in particular, condition (3) holds for all $a_i \in A_i$ such that $x_i(a_i) = 0$. Also, using (1) we can write

$$\sum_{a_i \in A_i} x_i^*(a_i) U_i(x^*) = \sum_{a_i \in A_i} x_i^*(a_i) U_i\left(a_i, x_{-i}^*\right) . \qquad (5)$$

If there is $a_i \in A_i$ such that $x_i^*(a_i) > 0$ and $U_i(x^*) > U_i\left(a_i, x_{-i}^*\right)$ then, using (4),

$$\sum_{a_i \in A_i} x_i^*(a_i) U_i(x^*) > \sum_{a_i \in A_i} x_i^*(a_i) U_i((a_i, x_{-i}^*))$$

in contradiction to (5). □

Corollary 1 *The strategy profile $x^* = (x_k^*)_{k \in N}$ is an equilibrium of the mixed extension of $\langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$ if and only if for all players $i \in N$ and for all $a_i \in A_i$,*

$$x_i^*(a_i) > 0 \quad \textit{implies } a_i \in \mathcal{B}_i(x_{-i}^*) .$$

According to the standard interpretation, a player's mixed strategy in a game G is an action, but in a different game, namely in the mixed extension of G. According to this interpretation, a mixed strategy is a deliberate choice of a player to use a random device. A mixed strategy equilibrium then is a profile of *independent* random devices, each of which is a best response to the others. Corollary 1 provides an alternative interpretation of a mixed strategy equilibrium. According to this interpretation, a player's mixed strategy represents the uncertainty in the minds of the other players concerning the player's action. In other words, a player's mixed strategy is interpreted not as a deliberate choice of the player but the belief, shared by all the other players, about the player's choice. That is, if $(x_k)_{k \in N}$ is a profile of mixed strategies, then x_i is the conjecture, shared by all the players other than i, about i's ultimate choice of action. Consequently, x_{-i} are the conjectures entertained by player i about his opponents' actions. According to this interpretation, Corollary 1 says that a mixed strategy equilibrium $(x_k^*)_{k \in N}$ is a profile of beliefs about each player's actions (entertained by the other players) according to which each player chooses an action that is a best response to his own beliefs.

The War of Attrition (cont.)

We have seen in Sect. "Analysis of Some Finite Games" that all the Nash equilibria of the war of attrition predict no real fight for the prey. We will now see that there is a mixed strategy equilibrium of the war of attrition that predicts a positive-length fight with probability one.

The players' action sets in the war of attrition are intervals of real numbers. A mixed strategy for player i in that game can be represented by a cumulative distribution function $F_i\colon [0, \infty] \to [0, 1]$. For each $t \in (0, \infty]$, $F_i(t)$ is the probability that player i gives up at or before t. We will look for a Nash equilibrium (F_1, F_2) that consists of two strictly increasing, differentiable cumulative distribution functions. The density of F_i is denoted by f_i. We will try to find an equilibrium at which each player is indifferent between all pure actions.

Consider player i. Given that his opponent is using mixed strategy F_j, $j \neq i$, if he chooses to give in at time t, then he will face a lottery according to which,

- With probability $1 - F_j(t)$, player i does not obtain the prey and gets a payoff of $-t$,
- With probability $F_j(t)$, player i obtains the prey at time t_j, where t_j is a random variable whose cumulative distribution function is $F_j(t_j)/F_j(t)$ (the distribution player j's surrender time, conditional on his having given in before t).

Therefore, the corresponding expected utility of choosing time t is

$$U_i(t, F_j) = (1 - F_j(t))(-t) + F_j(t)\int_0^t (v_i - t_j)\,\mathrm{d}\frac{F_j(t_j)}{F_j(t)}$$
$$= (1 - F_j(t))(-t) + \int_0^t (v_i - t_j)\mathrm{d}F_j(t_j)\,.$$

Since in the equilibrium we are looking for, player i is indifferent among all his actions, the above expression is independent of t. Namely, $U_i(t, F_j) \equiv c$. As a result, the derivative of the above utility with respect to t equals 0. Formally,

$$\frac{\partial U_i(t, F_j)}{\partial t} = tf_j(t) - (1 - F_j(t)) + (v_i - t)f_j(t)$$
$$= (1 - F_j(t)) + v_i f_j(t) = 0.$$

This is a differential equation whose general solution is

$$F_j(t) = 1 - K\mathrm{e}^{-\frac{t}{v_i}}\,.$$

If we want it to satisfy $F_j(0) = 0$, we obtain that $K = 1$. As a result, the distribution function is given by

$$F_j(t) = 1 - \mathrm{e}^{-\frac{t}{v_i}}\,.$$

Consequently, the equilibrium we are looking for is

$$(F_1(t), F_2(t)) = \left(1 - \mathrm{e}^{-\frac{t}{v_2}}, 1 - \mathrm{e}^{-\frac{t}{v_1}}\right)\,.$$

According to this equilibrium, for any t, the probability that there is a fight that lasts at least t is $(1 - F_1(t))(1 - F_2(t)) > 0$. Consequently, there is a fight with probability one. The introduction of mixed strategies allowed the concept of Nash equilibrium to be consistent with fights that last a positive length of time. However, the mixed strategy equilibrium has the following unfortunate property. If $v_1 < v_2$, then for all $t > 0$, $F_1(t) < F_2(t)$. In other words, it is more likely that the player with the highest willingness to fight for the prey gives up earlier than any given t, than that the player with the lowest willingness to fight gives in earlier than the same t. Therefore, in equilibrium it is more likely that the player with the lower willingness to fight wins the war than the other way around. In particular, the probability that player 1 gets the object is given by

$$\int_0^\infty F_2(t)\mathrm{d}F_1(t)$$

which can be checked to be equal to $\frac{v_2}{v_1+v_2} > 1/2$. In order to obtain the more intuitive result that the higher the willingness to fight for the prey, the higher is the probability to obtain it, we will need to model the war of attrition in yet a different way. We'll return to this when we introduce asymmetric information to the games.

Equilibrium in Beliefs

The mixed extension of the game $\langle N, (A_i)_{i\in N}, (u_i)_{i\in N}\rangle$ is constructed in two steps. First, we enlarge the set of actions available to each player by allowing him to choose any mixed strategy on his original action set. Second, since the action choices are now probability distributions over actions, we extend the players' original preferences to preferences over profiles of mixed strategies. We do so by evaluating each mixed strategy profile according to the expected value of the original utilities with respect to the probability distribution over action profiles induced by the mixed strategy.

The first step seems uncontroversial since it is certainly possible for players to use random devices. But the second step is somewhat problematic because, by evaluating mixed strategies according to the expected utility of the resulting lotteries, one is implicitly imposing on the players a certain kind of risk preferences. One may wonder what the implications would be if instead of extending the preferences by assuming that players are expected utility maximizers, we assume that players have more general preferences over profiles of mixed strategies. In particular, we would like to know if there is a suitable generalization of Corollary 1.

Let $G = \langle N, (A_i)_{i\in N}, (u_i)_{i\in N}\rangle$ be a finite game. We define the mixed extension of G as the strategic game $\langle N, (X_i)_{i\in N}, (U_i)_{i\in N}\rangle$ where, as in Sect. "Mixed Strategies", X_i is the set of probability distributions over the actions in A_i, for $i \in N$, but unlike there, the utility function $U_i\colon X \to \mathbb{R}^N$ is not necessarily a multilinear function of the probabilities, but a general continuous function of the mixed strategies. The only requirement on U_i is that for all profiles of degenerate mixed strategies $(a_k)_{k\in N}$, we have $U_i\left((a_k)_{k\in N}\right) = u_i\left((a_k)_{k\in N}\right)$. As before, a mixed strategy Nash equilibrium of $\langle N, (A_i)_{i\in N}, (u_i)_{i\in N}\rangle$ is a Nash equilibrium of its mixed extension $\langle N, (X_i)_{i\in N}, (U_i)_{i\in N}\rangle$. In other words, it is a list of mixed strategies $\left(x_k^*\right)_{k\in N}$ such that for all players $i \in N$ and for all of his mixed strategies x_i,

$$U_i(\left(x_k^*\right)_{k\in N}) \geq U_i(\left(x_i, x_{-i}^*\right)) .$$

Alternatively, $\left(x_k^*\right)_{k\in N}$ is a mixed strategy equilibrium if

$$x_i^* \in \mathcal{B}_i(x_{-i}^*) \quad \text{for all } i \in N .$$

Observation 1 It is important to note that two different actions of a player may be best responses to a given mixed strategy profile of the other players, and yet no probability mixture of the two actions will be a best response to the given mixed strategy profile. This will typically be the case when the function U_i is strictly convex in X_i, since strictly convex functions attain their maximum at boundary points.

Theorem 1 shows that Nash equilibria exist when the extended utility function U_i is concave in X_i. However, Observation 1 indicates that a Nash equilibrium may fail to exist when U_i is strictly convex in X_i. Indeed, take a game $G = \langle N, (A_i)_{i\in N}, (u_i)_{i\in N}\rangle$ with no pure strategy Nash equilibrium, like Matching Pennies, and consider its mixed extension $\Gamma = \langle N, (X_i)_{i\in N}, (U_i)_{i\in N}\rangle$ where for all players, their extended utility function is strictly convex. Then, for any player $i \in N$ and for any profile of mixed strategies x_{-i} of the other players, the set of i's best responses $\mathcal{B}_i(x_{-i})$ consists of only degenerate mixed strategies. Since G has no pure strategy Nash equilibrium, we conclude that Γ does not have a Nash equilibrium.

Observation 2 It is also important to note that, unlike in the standard expected utility case, a player's mixed strategy x_i^* may very well be a best response to some profile x_{-i}^* of the other players' mixed strategies and at the same time may assign positive probability to an action that (when regarded as a degenerate mixed strategy) is not a best response to x_{-i}^*. Formally, it may very well be the case that

$$U_i(\left(x_k^*\right)_{k\in N}) \geq U_i(\left(x_i, x_{-i}^*\right)) \quad \text{for all } x_i \in X_i$$

and yet

$$U_i(\left(a_i, x_{-i}^*\right)) < U_i(\left(x_k^*\right)_{k\in N}) \quad \text{for some } a_i \text{ such } \quad x_i^*(a_i) > 0 .$$

This will typically occur when the function U_i is strictly concave in X_i.

The definition of mixed strategy equilibrium requires from each strategy in the equilibrium profile that it be a best response to the other strategies. Corollary 1 stated that when preferences have the expected utility form, each mixed strategy in a mixed strategy equilibrium is also a probability mixture over best responses to the other strategies in the profile. This result allowed us to interpret a mixed strategy Nash equilibrium as a profile of beliefs, rather than as a profile of probability mixtures. As explained in Observation 2, however, when preferences over mixed strategies are not expected utility preferences, a mixture over best responses is not necessarily a best response. Therefore, Corollary 1 does not extend to the mixed extension where preferences are not of the expected utility form.

In this setup, however, one can still interpret a player's mixed strategy as a belief entertained by the other players about the actions chosen by that player. And a profile of such beliefs will be in equilibrium if the probability distribution over the player's actions that represents i's beliefs is obtained as a mixture of best responses of this player to his beliefs about the other players' actions. With this idea in mind, Crawford [8] defined the notion of an equilibrium in beliefs. Before we formally present his definition we need to introduce some notation.

Since when the extended utility functions U_i are concave in i's own strategy a best response to a given profile of the other players' strategies may be a non-degenerate mixed strategy, a mixture of best responses will typically be a mixture over non-degenerate mixed strategies. This mixture induces a probability distribution over actions in a natural way by reducing the compound mixture to a simple mixture. This induced probability distribution can be interpreted as a belief over the actions ultimately chosen. For example, in Matching Pennies, if player 1 believes that there is a probability of 1/2 that player 2 will choose the mixed strategy $(1/3, 2/3)$ and a probability of 1/2 that player 2 will choose the mixed strategy $(2/3, 1/3)$, then player 1 believes that player 2 will choose each one of his two actions with equal probability. More generally, if player i assigns probability p_k to the event that player j will choose mixed strategy $x^k \in X_j$, for $k = 1, \ldots, K$, then player i's beliefs about player j's actions are given by $\sum_{k=1}^K p_k x^k \in X_j$. That is, for each action $a_j \in A_j$ of player j, player i believes that player j will choose a_j with probability $\sum_{k=1}^K p_k x^k(a_j)$. For each set $T \subset X_i$ of mixed strategies, let $D[T] \subset X_i$ denote the set of probability distributions over i's actions that are induced by mixtures over elements of T. With this notation in hand, we can define the concept of equilibrium in beliefs.

Definition 5 Let $G = \langle N, (A_i)_{i\in N}, (u_i)_{i\in N}\rangle$ be a game. For each $i \in N$, let $\mathcal{B}_i \colon X \to X_i$ be the best response correspondence in the mixed extension of G. The profile of beliefs $(x^*_k)_{k\in N} \in \times_{k\in N}\Delta(A_k)$ is an *equilibrium in beliefs* if

$$x^*_i \in D\left[\mathcal{B}_i(x^*_{-i})\right] \quad \text{for all } i \in N\,.$$

An equilibrium in beliefs is a profile of beliefs $\left(x^*_k\right)_{k\in N}$. For each $i \in N$, x^*_i is the common belief of the players other than i about player i's choice of actions. In order for this profile of beliefs to be in equilibrium, we require that for each player $i \in N$ all the other players believe that i chooses a mixed strategy that is a best response to his beliefs, which are given by $\left(x^*_k\right)_{k\in N\setminus\{i\}}$, about the other players' choices of actions. In other words, x^*_i must be a convex combination of best responses of i to $\left(x^*_k\right)_{k\in N\setminus\{i\}}$.

Example 2 Consider again the mixed extension of Matching Pennies $\langle N, (X_i)_{i\in N}, (U_i)_{i\in N}\rangle$ where the set of players is $N = \{1,2\}$, the sets of mixed strategies are $X_1 = \{(p_H, p_T) \geq (0,0)\colon p_H + p_T = 1\}$ and $X_2 = \{(q_H, q_T) \geq (0,0)\colon q_H + q_T = 1\}$, and the utility functions are now given by $U_1((p_H,p_T),(q_H,q_T)) = (p_Hq_H)^2 + (p_Tq_T)^2 - p_Hq_T - p_Tq_H$ and $U_2((p_H,p_T),(q_H,q_T)) = (p_Hq_T)^2 + (p_Tq_H)^2 - p_Hq_H - p_Tq_T$. Since the utility functions are strictly convex in the players's own mixed strategies, the best response to any strategy of the opponent is a pure strategy. In particular, one can verify that

$$\mathcal{B}_1(q_H,q_T) = \begin{cases} (1,0) & \text{if } q_H > q_T \\ \{(1,0)\,,(0,1)\} & \text{if } q_H = q_T \\ (0,1) & \text{if } q_H < q_T \end{cases}$$

and

$$\mathcal{B}_2(p_H,p_T) = \begin{cases} (0,1) & \text{if } p_H > p_T \\ \{(1,0)\,,(0,1)\} & \text{if } p_H = p_T \\ (1,0) & \text{if } p_H < p_T\,. \end{cases}$$

It can also be verified that $((p^*_H,p^*_T),(q^*_H,q^*_T)) = ((1/2,1/2),(1/2,1/2))$ is an equilibrium in beliefs. Indeed, for both $i = 1,2$, $(1/2,1/2) \in X_i$ is a convex combination of $(1,0)$ and $(0,1)$, which are both in $\mathcal{B}_j\,(1/2,1/2)$, $j \neq i$. In this equilibrium,

1. Player 1 believes that player 2 will choose $(1,0)$ with probability 1/2, and $(0,1)$ with probability 1/2.
2. Therefore player 1 believes that player 2 will ultimately choose H and T, each with probability 1/2.
3. Given these beliefs, player 1's only best replies are $(1,0)$ and $(0,1)$, and
4. Player 2 believes that player 1 will choose each one with probability 1/2. As a result,
5. Player 2 believes that player 1 will ultimately choose H and T each with probability 1/2.
6. Given these beliefs, player 2's only best replies are $(1,0)$ and $(0,1)$, and
1. Player 1 believes that player 2 will choose $(1,0)$ with probability 1/2, and $(0,1)$ with probability 1/2.

The following result is a direct implication of the definition of an equilibrium in beliefs.

Proposition 2 (Crawford [8]) *Let $G = \langle N, (A_i)_{i\in N}, (u_i)_{i\in N}\rangle$ be a strategic game, and $\Gamma = \langle N, (X_i)_{i\in N}, (U_i)_{i\in N}\rangle$ be the mixed extension of G, where U_i is continuous but not necessarily multilinear.*

1. *Every mixed strategy Nash equilibrium of G is an equilibrium in beliefs.*
2. *If for all $i \in N$, U_i is quasiconcave in X_i, then every equilibrium in beliefs is a mixed strategy Nash equilibrium of G.*

Proof

1. Since $\mathcal{B}_i(x^*_{-i}) \subset D\left[\mathcal{B}_i(x^*_{-i})\right]$ for all $i \in N$, every Nash equilibrium is an equilibrium in beliefs.
2. When the utility function U_i is quasiconcave in i's mixed strategy, the set of best responses $\mathcal{B}_i(x^*_{-i})$ is a convex set. Therefore, $D\left[\mathcal{B}_i(x^*_{-i})\right] = \mathcal{B}_i(x^*_{-i})$, and any equilibrium in beliefs is a Nash equilibrium. □

Crawford [8] shows that although some games have no Nash equilibrium, every game has an equilibrium in beliefs.

Correlated Equilibrium

In the mixed extension of a game, players do not choose their actions directly, but rather choose probability distributions over their action sets according to which the actions are ultimately selected. The important feature about these probability distributions is that they represent independent random variables. The realization of one player's random variable does not give any information about the realization of the other players' random variables. There is nothing in the bare notion of equilibrium, however, that requires players' behavior to be independent. The basic feature of an equilibrium is that each player is best responding to the behavior of others, and that each player is free to choose any action in his action set. But one thing is that players can, if they so wish, change their behavior without the consent of others, and another different thing is to expect players' choices to be independent. Therefore, one could ask what would happen if the random devices players use to ultimately choose their actions were correlated. In that case, knowledge of the realization of one's random device would provide some partial information about the realization of the other players' random devices, and therefore of their choices. In equilibrium, a player should take this information into account. To illustrate this point, consider the game of Chicken.

		Driver 2	
		Slow Down	Speed up
Driver 1	Slow Down	$6, 6$	$2, 7$
	Speed up	$7, 2$	$0, 0$

This game has two pure-action Nash equilibria, and one equilibrium in mixed strategies. According to the mixed strategy Nash equilibrium, each player chooses Slow Down with probability 2/3 and Speed Up with probability 1/3. This mixed strategy equilibrium can be implemented by the following random device. Consider two random variables S_1 and S_2, whose joint distribution is given by the following table:

Driver 1 chooses his action as a function of the realization of S_1 and Driver 2 chooses his action as a function of the realization of S_2. (Neither player is informed of the realization of the other player's random variable.) In particular, Driver 1 chooses Slow Down if $S_1 = 1$ and Speed Up otherwise. Similarly, Driver 2 chooses Slow Down if $S_2 = 1$, and Speed Up otherwise. Note that according to this pattern of behavior, each player chooses to slow down with probability 2/3. But more importantly, since S_1 and S_2 are independent random variables, knowledge of the realization of one random variable does not give any information about the realization of the other one. Therefore, after Driver 1 learns the realization of S_1, he still believes that Driver 2 will choose Slow Down with probability 2/3 and consequently any choice is optimal, in particular the one described above. Similarly, after Driver 2 learns the realization of S_2, he still believes that Driver 1 will choose to slow down with probability 2/3, and his planned behavior continues to be optimal.

Static Games, Table 1
A random device

		S_2	
		1	2
S_1	1	4/9	2/9
	2	2/9	1/9

But what would happen if the joint distribution of S_1 and S_2, was not as presented in Table 1, but rather as follows?

		S_2	
		1	2
S_1	1	1/3	1/3
	2	1/3	0

To answer this question, assume that both players still choose their actions according to the previous pattern of behavior: Driver 1 chooses Slow Down if $S_1 = 1$, and Speed Up otherwise. The same holds for Driver 2. As a result, it is still true that each player chooses Slow Down with probability 2/3 and Speed Up with probability 1/3. However, since this time the conditioning random variables S_1 and S_2 are not independent, knowledge of the realization of S_1 affects the beliefs of Driver 1 about the probability with which Driver 2 chooses his actions. In particular, if $S_1 = 1$, Driver 1 updates his beliefs and assigns probability 1/2 to Driver 2 choosing either action, and consequently, Driver 1's only optimal action is Slow Down, which is precisely the choice dictated by the above pattern of behavior. Similarly, if $S_1 = 2$, Driver 1 should update his beliefs and assign probability one that Driver 2 will choose Slow Down. Consequently, Driver 1's best reply is to follow the above pattern of behavior and choose Speed Up. One can see that, given that the players know that the random variables S_1 and S_2 are correlated and they use this information accordingly, there is no incentive for either of them to deviate from the proposed pattern of behavior. Therefore, we can say that this pattern of behavior is an equilibrium. This notion of a correlated equilibrium was introduced in Aumann [2]. Before we give a formal definition we introduce the concept of a correlated strategy profile, which will play a central role not only in this section, but in the next one as well.

Definition 6 Let $G = \langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$ be a game. A *correlated strategy profile in* G consists of

- A finite probability space (Ω, π)
- For each player $i \in N$, a partition $\mathcal{P}_i$ of Ω into events of positive probability
- For each player $i \in N$, a function $\sigma_i : \Omega \to A_i$ which is measurable with respect to $\mathcal{P}_i$.

A correlated strategy profile is a description of what players do and know while playing the game G. The collection $\langle(\Omega, \pi), (\mathcal{P}_i)_{i \in N}\rangle$ represents the random devices used by the players to ultimately choose their actions. The underlying probability space that governs the players' random devices is (Ω, π). Ω is the set of states, and for each state ω, $\pi(\omega)$ is the probability that ω occurs. For each $i \in N$, the partition $\mathcal{P}_i$ represents player *i*'s information. Each element of the partition represents a different realization of the random device used by *i* to choose his action. Two states that belong to the same element of the partition $\mathcal{P}_i$ cannot be distinguished by *i*, while two states that belong to different partition cells can be distinguished by him. For each player $i, \sigma_i : \Omega \to A_i$ is the random variable that describes players *i*'s choice of action, $\sigma_i(\omega)$ being the action chosen by him at state ω. The measurability of σ_i with respect to $\mathcal{P}_i$ formalizes the requirement that the actions chosen by player *i* depend only on his information about the state of the world. Therefore, for any two states that belong to the same element of his partition, the actions chosen by *i* at those states must be the same. That is, for any $\omega, \omega' \in P \in \mathcal{P}_i$ we have $\sigma_i(\omega) = \sigma_i(\omega')$.

For example, the correlated strategy profile described earlier for the game of chicken can be formalized as $\langle(\Omega, \pi), (\mathcal{P}_i)_{i \in N}, (\sigma_i)_{i \in N}\rangle$, where $N = \{\text{I}, \text{II}\}$, and

- $\Omega = \{(1,1), (1,2), (2,1))\}$
- $\pi(\omega) = 1/3$ for all $\omega \in \Omega$
- $\mathcal{P}_{\text{I}} = \{\{(1,1), (1,2)\}, \{(2,1)\}\}$ and $\mathcal{P}_{\text{II}} = \{\{(1,1), (2,1)\}\}, \{(1,2)\}$
- $\sigma_{\text{I}}(\omega) = \begin{cases} \text{Slow Down} & \text{if } \omega \in \{(1,1),(1,2)\} \\ \text{Speed up} & \text{if } \omega \in \{(2,1)\} \end{cases}$
- $\sigma_{\text{II}}(\omega) = \begin{cases} \text{Slow Down} & \text{if } \omega \in \{(1,1),(2,1)\} \\ \text{Speed up} & \text{if } \omega \in \{(1,2)\}\,. \end{cases}$

According to this correlated strategy profile, there are three equally likely states, and the players can distinguish only one component of the state, namely the realization of their random variable. The players' actions are described by the functions σ_I and σ_{II} which depend only on the respective player's information.

In what follows we denote by $\sigma : \Omega \to A$ the function that associates with each $\omega \in \Omega$ the action profile induced by the strategies σ_k, for $k \in N$. That is, $\sigma = (\sigma_k)_{k \in N}$. Also, for any $i \in N$, $\sigma_{-i} = (\sigma_k)_{k \in N \setminus \{i\}}$ so that $\sigma = (\sigma_{-i}, \sigma_i)$. We are interested in correlated strategy profiles in which no player benefits by altering his behavior. These special profiles are introduced in the following definition.

Definition 7 Let $G = \langle N, (A_i)_{i \in N}, (u_i)_{i \in N}\rangle$ be a strategic game. A *correlated equilibrium of* G is a correlated strategy $\langle(\Omega, \pi), (\mathcal{P}_i)_{i \in N}, (\sigma_i)_{i \in N}\rangle$ such that for every $i \in N$ and every function $\tau_i : \Omega \to A_i$ that is measurable with respect to $\mathcal{P}_i$,

$$\sum_{\omega \in \Omega} \pi(\omega) u_i(\sigma_{-i}(\omega), \sigma_i(\omega)) \geq \sum_{\omega \in \Omega} \pi(\omega) u_i(\sigma_{-i}(\omega), \tau_i(\omega))\,. \quad (6)$$

The value $v_i = \sum_{\omega \in \Omega} \pi(\omega) u_i(\sigma_{-i}(\omega), \sigma_i(\omega))$ is player *i*'s correlated equilibrium payoff.

In a correlated strategy profile each player plans to condition his choice of action on the realization of a random variable, and the players' random variables may be correlated. A correlated strategy profile is a correlated equilibrium if no player can find an alternative way to condition his choice on the same random device, so that his expected utility is increased. Note that the player presumably chooses his strategy (his way to condition his actions on the outcomes of the random device) before he learns the realization of the device. Nonetheless, he evaluates the outcomes generated by the players' strategies by taking into account the precise correlation of the random devices on which outcomes players are conditioning their behavior.

Although strictly speaking mixed strategy Nash equilibria are not correlated equilibria, they do induce a correlated equilibrium distribution over action profiles. In order to state this claim, we need the following definition.

Definition 8 Let $\langle(\Omega, \pi), (\mathcal{P}_i, \sigma_i)_{i \in N}\rangle$ be a correlated strategy profile for G. Its induced probability distribution over action profiles is given by the function $p : A \to [0,1]$ defined by

$$p(a) = \pi(\{\omega \in \Omega : \sigma(\omega) = a\}) = \sum_{\{\omega \in \Omega : \sigma(\omega) = a\}} \pi(\omega) \quad \text{for all } a \in A\,.$$

Proposition 3 *Let* $G = \langle N, (A_i)_{i \in N}, (u_i)_{i \in N}\rangle$ *be a strategic game, and let* $x = (x_1, \cdots, x_n)$ *be a mixed strategy Nash equilibrium of G. Then, there is a correlated equilibrium* $\langle(\Omega, \pi), (\mathcal{P}_i)_{i \in N}, (\sigma_i)_{i \in N}\rangle$ *whose induced probability distribution over action profiles is the same as x's distribution.*

Proof Let $\langle(\Omega, \pi), (\mathcal{P}_i)_{i \in N}, (\sigma_i)_{i \in N}\rangle$ be defined as follows:

- $\Omega = A$
- $\pi(a) = \prod_{i \in N} x_i(a_i)$
- $\mathcal{P}_i(a) = \{b \in A : b_i = a_i\}$
- $\sigma_i(a) = a_i$.

We claim that $\langle(\Omega, \pi), (\mathcal{P}_i)_{i \in N}, (\sigma_i)_{i \in N}\rangle$ is a correlated equilibrium whose probability distribution is the same as x's distribution. Let $i \in N$. Since x is a mixed strategy Nash equilibrium, we know by Lemma 1 that for all $a_i \in A_i$

$$\begin{aligned} &\text{if } x_i(a_i) > 0 \quad \text{then } U_i(x) = U_i(a_i, x_{-i}) \\ &\text{if } x_i(a_i) = 0 \quad \text{then } U_i(x) \geq U_i(a_i, x_{-i}) \,. \end{aligned}$$

Consequently, for all $a_i \in A_i$

$$x_i(a_i) U_i(a_i, x_{-i}) \geq x_i(a_i) U_i(b_i, x_{-i}) \quad \text{for all } b_i \in A_i \,. \tag{7}$$

Now let $\tau_i \colon A \to A_i$ be a function that is measurable with respect to $\mathcal{P}_i$. Let $a_{-i} \in A_{-i}$ be a fixed profile of actions for players other than i. Letting $b_i = \tau_i(a_i, a_{-i})$, Eq. (7) implies that

$$x_i(a_i) U_i(a_i, x_{-i}) \geq x_i(a_i) U_i(\tau_i(a_i, a_{-i}), x_{-i}) \quad \text{for all } a_i \in A_i \,.$$

Adding over all $a_i \in A_i$,

$$\sum_{a_i \in A_i} x_i(a_i) U_i(a_i, x_{-i}) \geq \sum_{a_i \in A_i} x_i(a_i) U_i(\tau_i(a_i, a_{-i}), x_{-i}) \,.$$

Taking into account the definition of $U_i(a_i, x_{-i})$ and $U_i(\tau_i(a), x_{-i})$, and using the measurability of τ_i with respect to $\mathcal{P}_i$, we get

$$\begin{aligned} &\sum_{a_i \in A_i} x_i(a_i) \sum_{a_{-i} \in A_{-i}} \left(\prod_{j \in N \setminus \{i\}} x_j(a_j) \right) u_i(a_i, a_{-i}) \\ &\quad \geq \sum_{a_i \in A_i} x_i(a_i) \sum_{a_{-i} \in A_{-i}} \left(\prod_{j \in N \setminus \{i\}} x_j(a_j) \right) u_i(\tau_i(a), a_{-i}) \\ &\sum_{a \in A} \left(\prod_{j \in N} x_j(a_j) \right) u_i(a_i, a_{-i}) \\ &\quad \geq \sum_{a \in A} \left(\prod_{j \in N} x_j(a_j) \right) u_i(\tau_i(a), a_{-i}) \\ &\sum_{a \in A} \pi(a) u_i(a_i, a_{-i}) \geq \sum_{a \in A} \pi(a) u_i(\tau_i(a), a_{-i}) \\ &\sum_{a \in A} \pi(a) u_i(\sigma_i(a), \sigma_{-i}(a)) \geq \sum_{a \in A} \pi(a) u_i(\tau_{-i}(a), \sigma_{-i}(a)). \end{aligned}$$

This shows that $\langle(\Omega, \pi), (\mathcal{P}_i)_{i \in N}, (\sigma_i)_{i \in N}\rangle$ is a correlated equilibrium of G. Its induced probability distribution over action profiles is

$$\begin{aligned} p(a) &= \pi(\{b \in A \colon \sigma(b) = a\}) \\ &= \pi(\{b \in A \colon b = a\}) \\ &= \pi(a) \\ &= \prod_{i \in N} x_i(a_i) \,. \end{aligned}$$

□

Although a correlated strategy profile consists of a randomizing device used by the players, it turns out that the only feature of the device that determines whether or not the correlated strategy profile constitutes a correlated equilibrium is its induced probability distribution over the action profiles. This is shown by the next proposition.

Proposition 4 *Let $G = \langle N, (A_i)_{i \in N}, (u_i)_{i \in N}\rangle$ be a finite strategic game. Every correlated equilibrium probability distribution over action profiles can be obtained in a correlated equilibrium of G in which*

- $\Omega = A$
- $\mathcal{P}_i(a) = \{b \in A \colon b_i = a_i\}$.

Proof Let $\left\langle(\Omega', \pi'), (\mathcal{P}'_i, \sigma'_i)_{i \in N}\right\rangle$ be a correlated equilibrium of G. Consider the correlated strategy profile $\langle(\Omega, \pi), (\mathcal{P}_i, \sigma_i)_{i \in N}\rangle$ defined by

- $\Omega = A$
- $\pi(a) = \pi'(\{\omega \in \Omega \colon \sigma'(\omega) = a\})$ for each $a \in A$
- $\mathcal{P}_i(a) = \{b \in A \colon b_i = a_i\}$ for each $i \in N$ and for each $a \in A$
- $\sigma_i(a) = a_i$ for each $i \in N$.

It is clear that this correlated strategy profile induces the required distribution over action profiles. Indeed,

$$\begin{aligned} p(a) &= \pi(\{\omega \in \Omega \colon \sigma(\omega) = a\}) \\ &= \pi(\{a' \in A \colon a' = a\}) \\ &= \pi(a) \\ &= \pi'(\{\omega \in \Omega' \colon \sigma'(\omega) = a\}) \,. \end{aligned}$$

It remains to show that this profile is a correlated equilibrium. Take a function $\tau_i \colon A \to A_i$ that is measurable with respect to $\mathcal{P}_i$. Define $\tau'_i \colon \Omega' \to A_i$ by $\tau'_i(\omega) = \tau_i(\sigma'(\omega)) = \tau_i(\sigma'_{-i}(\omega), \sigma'_i(\omega))$. The function τ'_i is measurable with respect to $\mathcal{P}'_i$. Indeed, if $\omega' \in \mathcal{P}'_i(\omega)$, then $\sigma'_i(\omega') = \sigma'_i(\omega)$ by measurability of σ'_i with respect to $\mathcal{P}'_i$. Therefore, by definition of $\mathcal{P}_i$, $\mathcal{P}_i(\sigma'(\omega')) = \mathcal{P}_i(\sigma'(\omega))$, and both $\sigma'_i(\omega')$ and $\sigma'_i(\omega)$ belong to the same element of

$\mathcal{P}_i$. Since τ_i is measurable with respect to $\mathcal{P}_i$, we conclude that $\tau'_i(\omega') = \tau_i(\sigma'_i(\omega')) = \tau_i(\sigma'_i(\omega)) = \tau'_i(\omega)$.

Also,

$$\begin{aligned}\sum_{\omega\in\Omega} &\pi(\omega)u_i(\sigma_{-i}(\omega),\tau_i(\omega))\\ &= \sum_{a\in A}\pi(a)u_i(a_{-i},\tau_i(a))\\ &= \sum_{a\in A}\sum_{\{\omega\in\Omega' : \sigma'(\omega)=a\}}\pi'(\omega)u_i(\sigma'_{-i}(\omega),\tau_i(\sigma'(\omega)))\\ &= \sum_{a\in A}\sum_{\{\omega\in\Omega' : \sigma'(\omega)=a\}}\pi'(\omega)u_i(\sigma'_{-i}(\omega),\tau'_i(\omega))\\ &= \sum_{\omega\in\Omega'}\pi'(\omega)u_i(\sigma'_{-i}(\omega),\tau'_i(\omega))\,.\end{aligned}$$

In particular, for $\tau_i = \sigma_i$,

$$\sum_{\omega\in\Omega}\pi(\omega)u_i(\sigma_{-i}(\omega),\sigma_i(\omega)) = \sum_{\omega\in\Omega'}\pi'(\omega)u_i(\sigma'_{-i}(\omega),\sigma'_i(\omega))\,.$$

Since $\langle(\Omega',\pi'),(\mathcal{P}'_i,\sigma'_i)_{i\in N}\rangle$ is a correlated equilibrium,

$$\sum_{\omega\in\Omega'}\pi'(\omega)u_i(\sigma'_{-i}(\omega),\sigma'_i(\omega)) \geq \sum_{\omega\in\Omega'}\pi'(\omega)u_i(\sigma'_{-i}(\omega),\tau'_i(\omega))$$

and therefore

$$\sum_{\omega\in\Omega}\pi(\omega)u_i(\sigma_{-i}(\omega),\sigma_i(\omega)) \geq \sum_{\omega\in\Omega}\pi(\omega)u_i(\sigma_{-i}(\omega),\tau_i(\omega))\,.$$

□

Rationality, Correlated Equilibrium and Equilibrium in Beliefs

As mentioned earlier, Nash equilibrium and correlated equilibrium are two examples of what is known as solution concepts. Solution concepts assign to each game a pattern of behavior for the players in the game. The interpretation of these patterns of behavior is not always explicit, but it is fair to say that they are usually interpreted either as descriptions of what rational people do, or as prescriptions of what rational people should do. There is a growing literature that tries to connect various game theoretic solution concepts to the idea of rationality. Rationality is generally understood as the characteristic of a player who chooses an action that maximizes his preferences, given his information about the environment in which he acts. Part of the information a player has is represented by his beliefs about the behavior of other players, their beliefs about the behavior of other players, and so on. So when one speaks of the rationality of players, one needs to take into account their *epistemic state*. There is a formal framework which is appropriate for discussing the actions, knowledge, beliefs and rationality of players. Namely, the framework of a correlated strategy profile. As defined in Sect. "Correlated Equilibrium", a correlated strategy profile in a game G consists of

- A finite probability space (Ω,π)
- For each player $i \in N$ a partition $\mathcal{P}_i$ of Ω into events of positive probability
- For each player $i \in N$ a function $\sigma_i : \Omega \to A_i$ which is measurable with respect to $\mathcal{P}_i$.

For the present discussion we interpret a correlated strategy profile $\langle(\Omega,\pi),(\mathcal{P}_i)_{i\in N},(\sigma_i)_{i\in N}\rangle$ as a description of the players' behavior and beliefs, as observed by an outside observer. The set Ω is the set of possible states of the world and π is the prior probability on Ω shared by all the players. For each player $i \in N$, $\mathcal{P}_i$ is a partition of Ω that represents i's information. At state $\omega \in \Omega$, player i is informed not of the state that actually occurred, but of the element $\mathcal{P}_i(\omega)$ of his partition that contains ω. Player i then uses this information and his prior π to update his beliefs about the true state of the world. Finally, the function σ_i represents the actions taken by player i at each state. In particular, $\sigma_i(\omega)$ is the action chosen by i at state ω. Although a correlated equilibrium can be interpreted as a correlated strategy profile *prescribed* by a given solution concept (that of a correlated equilibrium), here we want to interpret a correlated strategy profile as a description of what players actually do and believe. Although players cannot freely choose their beliefs (in the same way as they cannot choose their preferences), they can choose their actions. Furthermore, they have no obligation to behave according to the specified correlated strategy profile. However, ultimately players do behave in a certain way and that behavior is what is represented by the given correlated strategy profile.

Once we fix a correlated strategy profile we can address the rationality of the players. Formally,

Definition 9 Player $i \in N$ is *Bayes rational* at $\omega \in \Omega$ if his expected payoff at ω, $E(u_i(\sigma)|\mathcal{P}_i)(\omega)$, is at least as large as the amount $E(u_i(\sigma_{-i}, a_i)|\mathcal{P}_i)(\omega)$ that he would have got had he chosen action $a_i \in A_i$ instead of $\sigma_i(\omega)$.

In other words, player i is rational at a given state of the world if the action $\sigma_i(\omega)$ he chooses at that state maximizes his expected utility given his information, $\mathcal{P}_i(\omega)$, and, in particular, given his beliefs about the actions of the other players.

As before, for any finite set T, let $\Delta(T)$ be the set of all probability distributions on T. The beliefs of player i about the actions of the other players are represented by his conjectures. A conjecture of i is a probability distribution $\psi_i \in \Delta(A_{-i})$ over the elements of A_{-i}. For any $j \neq i$, the marginal of ψ_i on A_j is the conjecture of i about j induced by ψ_i. Given a correlated strategy profile $\langle(\Omega, \pi), (\mathcal{P}_i)_{i\in N}, (\sigma_i)_{i\in N}\rangle$, one can determine the conjectures that each player is entertaining at each state of the world about the actions of the other players. These conjectures are given by the following definition.

Definition 10 Given a correlated strategy profile $\langle(\Omega, \pi), (\mathcal{P}_i, \sigma_i)_{i\in N}\rangle$, the conjectures of $i \in N$ about the other players' actions are given by the function $\phi_i \colon \Omega \to \Delta(A_{-i})$ defined by

$$\phi_i(\omega)(a_{-i}) = \frac{\pi\left[\{\omega' \in \mathcal{P}_i(\omega) \colon \sigma_{-i}(\omega') = a_{-i}\}\right]}{\pi\left[\mathcal{P}_i(\omega)\right]} .$$

For each ω, $\phi_i(\omega) \in \Delta(A_{-i})$ is the conjecture of i at ω. For $j \neq i$, the marginal of $\phi_i(\omega)$ on A_j is the conjecture of i at ω about j's actions.

Given a correlated strategy profile, we can speak about what each player knows. The object of knowledge are called *events*, which are the subsets of the set of states of the world Ω. We say that player i *knows* event $E \subset \Omega$ at state ω, if $P_i(\omega) \subset E$. That is, i knows E at ω if whatever state he deems possible at ω is in E.

The next result, proved by Aumann and Brandenburger [5], shows a remarkable relationship between the rationality of players and the concept of equilibrium in beliefs.

Theorem 2 *Fix a two-person game,* $G = \langle N, (A_i)_{i\in N}, (u_i)_{i\in N}\rangle$, *and let* $\langle(\Omega, \pi), (\mathcal{P}_i)_{i\in N}, (\sigma_i)_{i\in N}\rangle$ *be a correlated strategy profile for G. Let* $\psi_1 \in \Delta(A_1)$ *and* $\psi_2 \in \Delta(A_2)$ *be two conjectures, one about player 1's actions and the other about player 2's actions. Assume that at some state* $\omega \in \Omega$ *each player knows that the other is rational and that their conjectures at* ω *are* $(\phi_1(\omega), \phi_2(\omega)) = (\psi_2, \psi_1)$. *Then,* (ψ_1, ψ_2) *is an equilibrium in beliefs.*

Proof The fact that player i knows at ω that j's conjecture is ψ_i means that

$$\mathcal{P}_i(\omega) \subset \{\omega' \in \Omega \colon \phi_j(\omega')(a_i) = \psi_i(a_i) \quad \text{for all } a_i \in A_i\} .$$

Therefore

$$\phi_j(\omega)(a_i) = \psi_i(a_i) \quad \text{for all } a_i \in A_i . \tag{8}$$

Given Proposition 2 and Corollary 1, we need to show that if $\psi_i(a_i^*) > 0$, a_i^* is a best response to ψ_j, for $i, j = 1, 2$, $i \neq j$. For this purpose, assume that $\psi_i(a_i^*) > 0$ for some $a_i^* \in A_i$. Then, by definition of ϕ_j and (8), $\phi_j(\omega)(a_i^*) = \pi\left[\{\omega' \in \mathcal{P}_j(\omega) \colon \sigma_i(\omega') = a_i^*\}\right] > 0$. Consequently, there is $\omega' \in \mathcal{P}_j(\omega)$ such that $\sigma_i(\omega') = a_i^*$. Since player j knows at ω that player i is rational,

$$\omega' \in \mathcal{P}_j(\omega) \subset \left\{\omega'' \in \Omega \colon E\left[u_i(\sigma)|\mathcal{P}_i\right](\omega'') \geq E\left[u_i(\sigma_{-i}, a_i)|\mathcal{P}_i\right](\omega'') \quad \text{for all } a_i \in A_i\right\} .$$

Therefore,

$$E\left[u_i(\sigma)|\mathcal{P}_i\right](\omega') \geq E\left[u_i(\sigma_{-i}, a_i)|\mathcal{P}_i\right](\omega') \quad \text{for all } a_i \in A_i$$

and since $\sigma_i \colon \Omega \to A_i$ is measurable with respect to $\mathcal{P}_i$, $\sigma_i(\omega') = a_i^*$ is the action that player i chooses at all states in $\mathcal{P}_i(\omega')$. Then we can write

$$E\left[u_i(\sigma_{-i}, a_i^*)|\mathcal{P}_i\right](\omega') \geq E\left[u_i(\sigma_{-i}, a_i)|\mathcal{P}_i\right](\omega') \quad \text{for all } a_i \in A_i .$$

That is, for all $a_i \in A_i$

$$\sum_{\omega'' \in \mathcal{P}_i(\omega')} \frac{\pi(\omega'')}{\pi(\mathcal{P}_i(\omega'))} u_i(\sigma_{-i}(\omega''), a_i^*) \geq \sum_{\omega'' \in \mathcal{P}_i(\omega')} \frac{\pi(\omega'')}{\pi(\mathcal{P}_i(\omega'))} u_i(\sigma_{-i}(\omega''), a_i)$$

$$\sum_{a_j \in A_j} \sum_{\substack{\omega'' \in \mathcal{P}_i(\omega') \\ \sigma_j(\omega'') = a_j}} \frac{\pi(\omega'')}{\pi(\mathcal{P}_i(\omega'))} u_i(a_j, a_i^*) \geq \sum_{a_j \in A_j} \sum_{\substack{\omega'' \in \mathcal{P}_i(\omega') \\ \sigma_j(\omega'') = a_j}} \frac{\pi(\omega'')}{\pi(\mathcal{P}_i(\omega'))} u_i(a_j, a_i)$$

$$\sum_{a_j \in A_j} \frac{\pi\left[\{\omega'' \in \mathcal{P}_i(\omega') \colon \sigma_j(\omega'') = a_j\}\right]}{\pi(\mathcal{P}_i(\omega'))} u_i(a_j, a_i^*)$$
$$\geq \sum_{a_j \in A_j} \frac{\pi\left[\{\omega'' \in \mathcal{P}_i(\omega') \colon \sigma_j(\omega'') = a_j\}\right]}{\pi(\mathcal{P}_i(\omega'))} u_i(a_j, a_i)$$
$$\sum_{a_j \in A_j} \phi_i(\omega')(a_j) u_i(a_j, a_i^*)$$
$$\geq \sum_{a_j \in A_j} \phi_i(\omega')(a_j) u_i(a_j, a_i)\,. \tag{9}$$

Since $\omega' \in \mathcal{P}_j(\omega)$ and player j knows at ω that i's conjecture is ψ_j, then

$$\omega' \in \mathcal{P}_j(\omega) \subset \left\{\omega'' \in \Omega \colon \phi_i(\omega'')(a_j) = \psi_j(a_j) \text{ for all } a_j \in A_j\right\}.$$

Therefore $\phi_i(\omega')(a_j) = \psi_j(a_j)$ for all $a_j \in A_j$, or $\phi_i(\omega') = \psi_j$. That is, i's conjecture at ω' about j's actions is ψ_j. Consequently, substituting in (9),

$$\sum_{a_j \in A_j} \psi_j(a_j) u_i(a_j, a_i^*) \geq \sum_{a_j \in A_j} \psi_j(a_j) u_i(a_j, a_i) \quad \forall a_i \in A_i\,.$$

That is, a_i^* is a best response to player i's beliefs about j's actions. □

The only assumptions required by Theorem 2 is that players know they are rational, and that they know each other's conjectures. In a correlated strategy profile for a two-player game, there is only one player entertaining a conjecture about the actions of player 1, namely, player 2. Similarly, player 1 is the only one who entertains a conjecture about the actions of player 2. In an n-person game, with $n > 2$, for each player, there is more than one player entertaining a conjecture about his actions. Therefore, since an equilibrium in beliefs consists of a profile of beliefs, each of which is shared by $n-1$ players, a generalization of Theorem 2 would require the players' beliefs about player i's actions, for $i \in N$, to be identical. In order to obtain these common beliefs it is not sufficient to assume that players know each other's conjectures. One need to strengthen this assumption. Also, in an equilibrium in beliefs, the common belief about player i's actions assigns positive probability only to best responses to i's conjectures about the choices of the other players. Furthermore, i's conjectures about the other players' choices is the product of his beliefs about each of the other players. In other words, an equilibrium in beliefs implicitly assumes that players believe that the other players' choices are independent. Aumann and Brandenburger [5] show that one way to obtain common conjectures and, simultaneously, that players believe that the other players act independently, is to assume that players' conjectures are commonly known. This surprising and deep result is stated in the next theorem.

Theorem 3 *Let $G = \langle N, (A_i)_{i \in N}, (u_i)_{i \in N}\rangle$ be a strategic game, and let $\langle(\Omega, \pi), (\mathcal{P}_i)_{i \in N}, (\sigma_i)_{i \in N}\rangle$ be a correlated strategy profile for G. Also let $(\psi_i)_{i \in N} \in \times_{i \in N} \Delta(A_{-i})$ be a profile of conjectures, one for each player. Assume that at some state $\omega \in \Omega$ each player knows that the others are rational. Further, assume that at ω their conjectures are commonly known to be $(\psi_i)_{i \in N}$. Then, for each j, all the conjectures ψ_i of players i other than j, induce the same belief $\varphi_j \in \Delta(A_j)$ about j's actions, and the resulting profile of beliefs, $(\varphi_i)_{i \in N}$, is an equilibrium in beliefs.*

Rationality and Correlated Equilibrium

The previous result shows a surprising relationship between the players' rationality and the concept of equilibrium in beliefs. If at some state of the world players know that everybody is rational, and if their conjectures are commonly known at that state, then their beliefs about each player's actions are in equilibrium. It is not that their actions constitute an equilibrium, but that their beliefs do. The question that naturally arises is: are there any epistemic conditions on the players that would induce them to play according to equilibrium? To answer this we turn to Aumann [3], where it is stated that if players are rational at every state, then their behavior constitutes a correlated equilibrium. Therefore, in order to obtain an equilibrium behavior, a sufficient condition is not that players be rational, or that they know that they are rational at some particular state, but that their rationality be common knowledge. And if it is common knowledge that all players are rational, then their behavior is not necessarily a Nash equilibrium, but a correlated equilibrium.

Theorem 4 *Let G be a strategic game, and let $\langle(\Omega, \pi), (\mathcal{P}_i)_{i \in N}, (\sigma_i)_{i \in N}\rangle$ be a correlated strategy profile for G. If each player is rational at each state of the world, then $\langle(\Omega, \pi), (\mathcal{P}_i)_{i \in N}, (\sigma_i)_{i \in N}\rangle$ is a correlated equilibrium.*

Proof Let $\tau_i \colon \Omega \to A_i$ be a function that is measurable with respect to $\mathcal{P}_i$. Since i is Bayes rational at ω

$$E(u_i(\sigma)|\mathcal{P}_i)(\omega) \geq E(u_i(\sigma_{-i}, a_i)|\mathcal{P}_i)(\omega) \quad \forall a_i \in A_i\,.$$

That is,

$$\sum_{\omega' \in \mathcal{P}_i(\omega)} \frac{\pi(\omega')}{\pi(\mathcal{P}_i(\omega))} u_i(\sigma_{-i}(\omega'), \sigma_i(\omega')) \geq \sum_{\omega' \in \mathcal{P}_i(\omega)} \frac{\pi(\omega')}{\pi(\mathcal{P}_i(\omega))} u_i(\sigma_{-i}(\omega'), a_i) \quad \forall a_i \in A_i \,.$$

In particular, for $a_i = \tau(\omega) = \tau(\omega')$ for all $\omega' \in \mathcal{P}_i(\omega)$,

$$\sum_{\omega' \in \mathcal{P}_i(\omega)} \frac{\pi(\omega')}{\pi(\mathcal{P}_i(\omega))} u_i(\sigma_{-i}(\omega'), \sigma_i(\omega')) \geq \sum_{\omega' \in \mathcal{P}_i(\omega)} \frac{\pi(\omega')}{\pi(\mathcal{P}_i(\omega))} u_i(\sigma_{-i}(\omega'), \tau(\omega')) \,.$$

Multiplying both sides by $\pi(\omega)$ and adding over all the elements of $\mathcal{P}_i$ we get

$$\sum_{\omega \in \Omega} \pi(\omega) u_i(\sigma_{-i}(\omega), \sigma_i(\omega)) \geq \sum_{\omega \in \Omega} \pi(\omega) u_i(\sigma_{-i}(\omega), \tau(\omega)) \,.$$

□

Bayesian Games

Thus far, we have considered static games, which are objects of the form $\langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$. Although these games have many applications, they are not readily suitable for the analysis of situations involving asymmetric information. Indeed, an implicit assumption behind the definition of a static game is that all players have the same information about the relevant aspects of the situation. In particular, all players have the same information about the sets of actions and preferences of all players. A static game seems suitable to model strategic interactions like the prisoner's dilemma, rock scissors paper, and even chess. At the time they choose their actions, all the players have exactly the same information. There might be what is called strategic uncertainty, namely, uncertainty about what the players will do, but there is no uncertainty about the rules of the game and about the preferences of the players. But how would one translate a game of cards like bridge or poker into a static game? In a game of cards, at the time of choosing his actions, each player knows the cards he holds in his hand, but does not know the cards of his opponents. He only has a belief about the cards held by his opponents. In order to make a sound choice, a player will try to predict the actions of his opponents, but for this it is crucial to use his beliefs about the cards they hold. For the same reason, his opponents should use their beliefs about their own opponents' cards in order to make a sound choice. Thus, the beliefs about the cards held by each player should be part of a description of a game with asymmetric information. Further, in order to predict his opponents' actions, a player also needs to assess his opponents' beliefs about his own cards. This seems to induce an intractable infinite regress of beliefs, and beliefs about beliefs. Harsanyi [12] provided the basic structure to describe and analyze strategic situations where players are asymmetrically informed. This structure is called a Bayesian game.

Definition 11 A *Bayesian Game* is a system $\langle N, (\Omega, \mu), (A_i, \mathcal{P}_i, u_i)_{i \in N} \rangle$ where

- N is the set of players
- Ω is the set of states of nature
- μ is the players' common prior belief (a probability measure over the set of states)
- A_i is player i's set of actions
- $\mathcal{P}_i$ is player i's information partition (a partition of Ω into sets of positive measure). Each element of the partition is referred to as a player's type.
- $u_i\colon \times_{i \in N} A_i \times \Omega$ is player i's Bernoulli utility function (a function over pairs (a, ω) where $a \in A$ and $\omega \in \Omega$, the expected value of which represents the player's preferences among lotteries over the set of such pairs).

The interpretation of a Bayesian game is as follows. The basic uncertainty is represented by the probability space (Ω, μ) of all states of nature and the prior probability over them. Each state represents a realization of all the parametric uncertainty of the model. For instance, in a game of cards, each state represents each of the possible card deals. The information of player $i \in N$ is represented by his information partition $\mathcal{P}_i$. While states in the same element of the partition cannot be distinguished by the player, he can distinguish between states that belong to different partition cells. In a game of cards, for instance, each partition cell represents a particular set of cards dealt to the player. The probability measure μ represents the players' prior belief about the state of nature. This prior belief will be used along with the information obtained by each player to form beliefs about the other players' information. The set of actions of player i is A_i. Note that there is no loss of generality in assuming that this set does not depend on the state of nature. One can always add unavailable actions and assign them intolerable disutility. Finally, u_i is the payoff function that associates to each state of nature and action profile a utility level. Note that since the state of the world is unknown to the player at the time of making his choice, a player faces a lottery for any given action profile. The assumption is that the player evaluates this lottery according to the expected value of u_i with respect to that lottery.

Let $\langle N, (\Omega, \mu), (A_i, \mathcal{P}_i, u_i)_{i \in N}\rangle$ be a Bayesian game. A strategy for player $i \in N$ is a function $\sigma_i \colon \Omega \to A_i$ that is measurable with respect to $\mathcal{P}_i$. We denote the set of strategies for player i by $\mathcal{B}_i$. That is, $\mathcal{B}_i = \{\sigma_i \colon \Omega \to A_i \colon \sigma_i \text{ is measurable w.r.t. } \mathcal{P}_i\}$. The interpretation of a strategy in a Bayesian game is the usual one. For each state of nature $\omega \in \Omega, \sigma_i(\omega)$ is the action chosen by player i at ω. The measurability requirement imposes that player i's actions depend only on his information. If player i cannot distinguish between two states of nature, then he must choose the same action at both states. Player i evaluates a profile $\sigma \colon \Omega \to A$ of strategies according to the expected value of u_i with respect to μ.

In order to define an equilibrium notion for Bayesian games we follow the same idea used for the definition of a mixed strategy equilibrium. Namely, we translate the Bayesian game into a standard game, and then define an equilibrium of the Bayesian game as the Nash equilibirum of the induced game.

Definition 12 A *Bayesian equilibrium* of a Bayesian game $\langle N, (\Omega, \mu), (A_i, \mathcal{P}_i, u_i)_{i \in N}\rangle$ is a Nash equilibrium of the strategic game: $\langle N, (\mathcal{B}_i)_{i \in N}, (U_i)_{i \in N}\rangle$ where for each profile $\sigma \colon \Omega \to A$ of strategies, $U_i(\sigma) = E_\mu[u_i(\sigma(\omega), \omega)]$ is i's expected utility with respect to μ.

A Bayesian equilibrium of a Bayesian game is a Nash equilibrium of a properly defined static game. As such, conditions for its existence can be derived from Theorem 1. However, in many situations one is interested in particular kinds of equilibria. Specifically, in the analysis of auctions or of the war of attrition, one is often interested in efficient outcomes. In a single object auction, efficient outcomes are characterized by the fact that in equilibrium the object is allocated to the buyer who values it most. According to many standard auction rules, the object goes to the highest bidder. Therefore, in such auctions, to guarantee an efficient outcome, one would need a monotone equilibrium, namely, one in which bidders bids are higher the higher their valuations for the object are. Athey [1] shows conditions under which a Bayesian equilibrium exists where strategies are non-decreasing. The crucial conditions are that the players' types can be represented by a one-dimensional variable, and that, fixing a nondecreasing strategy for each of a player's opponents, this player's expected payoffs satisfies a single-crossing property. This single-crossing property roughly says that if a high action is preferred to a low action for a given type t, then the same must be true for all types higher than t. McAdams [16] extended Athey's result to the case where types and actions are multidimensional and partially ordered.

The Asymmetric Information Version of the War of Attrition

We have seen that, when applied to the war of attrition, as modeled by a standard strategic game or by its mixed extension, the notion of Nash equilibrium does not yield a satisfactory prediction.[1] In the former case all the equilibria involve no fight, and in the latter case the equilibrium dictates a more aggressive behavior to the player who values the contested object less. In what follows, we analyze the war of attrition as a Bayesian game. That is, we assume that the players are ex-ante symmetric but they have private information about their value for the contested object.

A Bayesian game that represents the war of attrition is given by $\langle N, \Omega, (A_i, \mu_i, \mathcal{P}_i, u_i)_{i \in N}\rangle$ where

- $N = \{1, 2\}$
- $\Omega = [0, \infty)^2 = \{(v_1, v_2) \colon 0 \le v_i < \infty, i = 1, 2\}$
- $A_i = [0, \infty)$ for $i = 1, 2$
- $\mathcal{P}_i(\hat{v}_1, \hat{v}_2) = \{(v_1, v_2) \in \Omega \colon v_i = \hat{v}_i\}$ for $i = 1, 2$
- $\mu((v_1, v_2) \le (\hat{v}_1, \hat{v}_2)) = F(\hat{v}_1) \times F(\hat{v}_2)$
- $u_i((a_1, a_2), (v_1, v_2)) = \begin{cases} -a_i & \text{if } a_i \le a_j \\ v_i - a_j & \text{if } a_i > a_j \, . \end{cases}$

Here the set of types of player i, for $i = 1, 2$, is represented by the player's willingness to fight, v_i. The players' willingness to fight are drawn independently from the same distribution F. A state of the world is, therefore, a realization (v_1, v_2) of the players' types, and at that state, each player is informed only of his type. Finally, the utility of a player is his valuation for the prey, if he obtains it, net of the time spent fighting for it. We are interested in a symmetric equilibrium in which both players use a symmetric, strictly increasing strategy $\beta \colon [0, \infty) \to [0, \infty)$, where $\beta(v)$ is the time at which a player with willingness to fight v is dictated by the equilibrium to give up. Such an equilibrium would imply that types who value the prey more, are willing to fight more. Further, the probability of observing a fight in equilibrium would not be 0 (in fact, it would be 1.)

It turns out that a symmetric equilibrium strategy is given by

$$\beta(v) = \int_0^v \frac{x f(x)}{1 - F(x)} \mathrm{d}x ,$$

[1] The war of attrition was analyzed in Maynard Smith [14]. For an analysis of the asymmetric information version of the war of attrition, see Krishna and Morgan [13].

where f denotes the derivative of F. To see this, assume that player j behaves according to β and that player i chooses to give up at t. Letting z be the type such $\beta(z) = t$, the expected utility of player i from choosing t is

$$U(v_i, z) = \int_0^z (v_i - \beta(y)) f(y) \mathrm{d}y - \beta(z)(1 - F(z)) \,.$$

Taking derivatives with respect to z, and using the fact that $\beta'(z) = z f(z)/[1 - F(z)]$ we obtain

$$\begin{aligned}\frac{\partial U}{\partial z}(v_i, z) &= v_i f(z) - \beta'(z)(1 - F(z)) \\ &= (v_i - z) f(z),\end{aligned}$$

which is positive for $z < v_i$, and negative for $z > v_i$. As a result, the expected utility of player i with willingness to pay v_i is maximized at $z = v_i$, which implies that the optimal choice is $\beta(v_i)$.

Thus, modeling the war of attrition as an asymmetric game has allowed us to find an equilibrium in which players with higher willingness to fight fight more, and there is a non-negligible probability of observing a fight.

Evolutionary Stable Strategies

The notion of the Nash equilibrium concept involves players choosing actions that maximize their payoffs given the choices of the other players. The usual interpretation of a Nash equilibrium is as a pattern of behavior that rational players *should* adopt. However, Nash equilibria are sometimes interpreted more descriptively as patterns of behavior that rational players *do* adopt. Certainly, rationality of players is neither a necessary condition nor a sufficient one for players to play a Nash equilibrium. The relationship between rationality and the various solution concepts is not apparent and has been the focus of an extensive literature (see, for example, [3,4,5,7]). Nonetheless, the notion of a Nash equilibrium evokes the idea of players consciously making choices with the deliberate objective of maximizing their payoffs. It is therefore quite remarkable that a concept almost identical to that of Nash equilibrium has emerged from the biology literature. This concept describes a population equilibrium where unconscious organisms are programmed to choose actions with no deliberate aim. In this equilibrium, members of the population meet at random over and over again to interact. At each interaction, these players act in a pre-programmed way and the result of their actions is a gain in biological fitness. Fitness is a concept related to the reproductive value or survival capacity of an organism. In a temporary equilibrium, the fitness gains are such that the proportions of individuals that choose each one of the possible actions remain constant. However, this temporary equilibrium may be disturbed by the appearance of a mutation, which is a new kind of behavior. This mutation may upset the temporary equilibrium if its fitness gains are such that the new behavior spreads over the population. Alternatively, if the fitness gains of the original population outweigh those of the mutation, then the new behavior will fail to propagate and will eventually disappear. In a population equilibrium, the interaction of any mutant with the whole population awards the mutant insufficient fitness gains, and as a result the mutants disappear. The notion of a population equilibrium is formalized by means of the concept of an evolutionary stable strategy, introduced by Maynard Smith and Price [15].

In what follows we restrict our attention to symmetric two-player games. So let $G = \langle \{1,2\}, \{A_1, A_2\}, \{u_1, u_2\} \rangle$ be a game such that $A_1 = A_2 = \overline{A}$, and such that for all $a, b \in \overline{A}, u_1(a,b) = u_2(b,a)$. An evolutionary stable strategy is an action in $\overline{A}$ such that if all members of the population were to choose that action, no sufficiently small proportion of mutants choosing an alternative action would succeed in invading the population. Alternatively, an evolutionary stable strategy is an action in $\overline{A}$ such that if all the members of the population were to choose that action, the population would reject all sufficiently small mutations involving a different action.

More specifically, suppose that all members of the population are programmed to choose $a \in \overline{A}$, and then a proportion ε of the population mutates and adopts action $b \in \overline{A}$. In that case, the probability that a given member of the population meets a mutant is ε, while the probability of meeting a member that plays a is $1 - \varepsilon$. Therefore, the mutation will not propagate and will vanish if the expected payoff of a mutant is less than the expected payoff of a member of the majority. Otherwise it will propagate. This leads to the following definition.

Definition 13 An action $a \in \overline{A}$ is an evolutionary stable strategy of G if there is an $\bar{\varepsilon} \in (0,1)$ such that for all $\varepsilon \in (0, \bar{\varepsilon})$, and for all $b \in \overline{A}$

$$(1-\varepsilon)u_1(a,a) + \varepsilon u_1(a,b) > (1-\varepsilon)u_1(b,a) + \varepsilon u_1(b,b) \,. \tag{10}$$

The following result shows that the concept of an evolutionary stable strategy is very close to the notion of a Nash equilibrium.

Proposition 5 *If $a \in \overline{A}$ is an evolutionary stable strategy of G, then (a,a) is a Nash equilibrium. And if (a,a) is* *a strict Nash equilibrium then a is an evolutionary stable strategy.*

Proof If $u_1(a,a) > u_1(b,a)$ for all $b \in \overline{A} \setminus \{b\}$, then inequality (10) holds for all sufficiently small $\varepsilon > 0$. If $u_1(b,a) > u_1(a,a)$ for some $b \in \overline{A}$, the reverse inequality holds for all sufficiently small ε. □

Future Directions

Static games have been shown to be a useful framework for analyzing and understanding many situations that involve strategic interaction. At present, a large body of literature is available that develops various solution concepts, some of which are refinements of Nash equilibrium and some of which are coarsenings of it. Nonetheless, several areas for future research remain. One is the application of the theory to particular games to better understand the situations they model, for example auctions. In many markets trade is conducted by auctions of one kind or another, including markets for small domestic products as well as some centralized electricity markets where generators and distributors buy and sell electric power on a daily basis. Also, auctions are used to allocate large amounts of valuable spectrum among telecommunication companies. It would be interesting to calculate the equilibria of many real life auctions. Simultaneously, future research should also focus on the design of auctions whose equilibria have certain desirable properties.

Another future direction would be to empirically and experimentally test the theory. The various equilibrium concepts predict certain kinds of behavior in certain games. Our confidence in the predictive and explanatory power of the theory depends on its performance in the field and in the laboratory. Moreover, the experimental and empirical results should provide valuable feedback for further development of the theory. Although some valuable experimental and empirical tests have already been performed (see [17,19,22,25] to name a few), the empirical aspect of game theory in general, and of static games in particular, remains underdeveloped.

Bibliography

1. Athey S (2001) Single crossing properties and the existence of pure strategy equilibria in games of incomplete information. Econometrica 69:861–889
2. Aumann RJ (1974) Subjectivity and correlation in randomized strategies. J Math Econ 1:67–96
3. Aumann RJ (1987) Correlated equilibrium as an expression of bayesian rationality. Econometrica 55:1–18
4. Aumann RJ (1995) Backward induction and common knowledge of rationality. Games Econ Behav 8:6–19
5. Aumann RJ, Brandenburger A (1995) Epistemic conditions for Nash equilibrium. Econometrica 63:1161–1180
6. Binmore K (2007) Playing for Real. Oxford University Press, New York
7. Brandenburger A, Dekel E (1987) Rationalizability and correlated equilibria. Econometrica 55:1391–1402
8. Crawford V (1990) Equilibrium without independence. J Econ Theory 50:127–154
9. Fudenberg D, Tirole J (1991) Game Theory. MIT Press, Cambridge
10. Geanakoplos J (2003) Nash and Walras equilibrium via Brouwer. Econ Theory 21:585–603
11. Glicksberg IL (1952) A further generalization of the Kakutani fixed point theorem, with application to Nash equilibrium points. Proc Am Math Soc 3:170–174
12. Harsanyi J (1967) Games with incomplete information played by 'Bayesian' players. Manag Sci Part i 14:159–82; Part ii 14:320–34; Part iii 14:486–502
13. Krishna V, Morgan J (1997) An analysis of the war of attrition and the all-pay auction. J Econ Theory 72:343–362
14. Smith MJ (1974) The theory of games and the evolution of animal conflicts. J Theor Biol 47:209–221
15. Smith MJ, Price GR (1973) The logic of animal conflict. Nature 246:15–18
16. McAdams D (2003) Isotone equilibrium in games of incomplete information. Econometrica 71:1191–1214
17. McKelvey RD, Palfrey TR (1992) An experimental study of the centipede game. Econometrica 60:803–836
18. Nash JF (1950) Equilibrium points in n-person games. Proc Natl Acad Sci USA 36:48–49
19. O'Neill B (1987) Nonmetric test of the minimax theory of two-person zero-sum games. Proc Natl Acad Sci 84:2106–2109
20. Osborne MJ (2004) An Introduction to Game Theory. Oxford University Press, New York
21. Osborne MJ, Rubinstein A (1994) A Course in Game Theory. MIT Press, Cambridge
22. Palacios-Huerta I (2003) Professionals play minimax. Rev Econ Stud 70:395–415
23. von Neumann J (1928) Zur Theorie der Gesellschaftsspiele. Math Ann 100:295–320
24. von Neumann J (1959) On the theory of games of strategy. In: Tucker AW, Luce RD (eds) Contributions to the Theory of Games, vol IV. Princeton University Press, Princeton
25. Walker M, Wooders J (2001) Minimax play at Wimbledon. Am Econ Rev 91:1521–1538

Statistical Applications of Wavelets

Sofia Olhede
Department of Statistical Science,
University College London, London, UK

Article Outline

Glossary

Locally stationary process A stochastic process that when sampled in time, its sampled points have joint characteristics in a short time window that only depend on the difference between the sampled time points, and not the global time point of the samples within the time window.

Multiresolution analysis The analysis of a phenomenon of interest over different temporal or spatial scales and locations.

Shrinkage A method of estimation where the initial estimator is decreased by multiplication with a factor of magnitude less than or equal to one. If the object being estimated is small, then a reduced variance will arise, that sets off the increased bias in estimation.

Sparsity Sparsity in a vector corresponds to the support of the vector being limited. The sparsity can either be strict, i. e. the vector is perfectly supported in a small subset of all its entries, or the magnitudes of the entries decay sufficiently rapidly, if ordered in magnitude.

Stationary process A stochastic process that when sampled in time, its sampled points have joint characteristics that only depend on the difference between the sampled time points, and not the global time point of the samples.

Thresholding A method of estimation where the initial estimator is set to zero if it does not exceed a given threshold. Thresholding is a special case of shrinkage.

Wavelet and scaling functions The functions that the wavelet and scaling filters converge to in increasing scale, when using discrete wavelet filters. These are not for any finite scale equivalent to the wavelet and scaling filters.

Wavelet and scaling filters The high and low pass digital filters that are used to calculate the discrete wavelet transform of a given sequence or vector. These are a pair of quadrature mirror filters that satisfy the perfect reconstruction property.

Definition of the Subject

A wavelet is a function essentially supported over a small time-window, that is oscillatory, with a local period that ranges over a limited sets of values. A wavelet function can be considered jointly local in time and frequency. Wavelets have become an integral part of statistical analysis of structured phenomena. Their utility for inference is based on several of their important properties, namely compression, separation of multiscale features, and differencing of a given set of moments. The statistical analysis of processes using wavelets presents a more faceted view of the phenomenon of interest than traditional methods. The utility of this in many areas is unquestionable, and statistical applications of wavelets include areas as disparate as climatology, econometrics, finance, geophysics, mathematics, as well as signal and image processing.

Introduction

The field of wavelets in statistics can roughly be divided into two areas, namely the usage of continuous wavelets, and the usage of discrete wavelets. Continuous wavelets are defined at a continuum of locations, whilst discrete wavelet filters are sequences defined only at a countable number of locations and scales. These sequences are not strictly speaking wavelet functions, but converge in increasing scale to such functions. Unfortunately in the literature both functions and filters are referred to as 'wavelets'.

Continuous wavelets are applied to the characterization of random processes, but are not usually used for actual signal estimation, as the transform does not possess an exact discretely implementable inverse. Continuous wavelets do not necessarily satisfy the constraint of compactness in time, and this permits more freedom in their design. A third class of wavelets that often receive separate treatment are complex-valued discrete wavelet filters. These can be thought of as bridging the gap between continuous and discrete representations. Complex wavelets are often chosen to be redundant, and have interpretable magnitude and phase, see for example the review of these in [48].

The statistical representation of structured phenomena using decompositions was first introduced by [47] who analyzed time series using Fourier decompositions. This analysis method corresponds to decomposing a sample into cosines and sines to determine the power associated with each frequency or mode. The interpretation of the observed weight attached to each frequency depends on the model posited for the observed data. If the process is assumed to be *stationary*, then the observed weight can be shown to converge in expectation to the true weight, namely the Fourier transform of the autocovariance of the time series. By calculating the full set of weights across frequencies, we may obtain a picture of the frequency content of the process. Even for processes that are not *stationary*, decomposing the set of observations into sinusoids allows us to visualize the frequency content of the signal. As long as the process is *harmonizable*, see [36], such decompositions are suitable.

Unfortunately whilst the decomposition of a time-varying process into sines and cosines is interpretable in a global sense, and can be shown to possess some suitable properties, it is not as easily interpreted as might be the case. Many processes are generated by mechanisms that fundamentally alter over time. It is therefore suitable to use decomposition functions that unlike sines are limited in time, but still correspond to oscillatory functions. A popular choice of such functions is wavelets.

A wavelet is a time-limited oscillation, or a "little wave". The first family was constructed by Alfred Haar in 1910, see [26]. The subject lay dormant for many years, until the mid 1980s when their usage in geophysics revitalized interest in the set of decompositions, by Morlet, and one of the groundbreaking new articles corresponds to [24]. A great theoretical leap forward was the construction of discrete wavelet filters, see [10]. Most of the statistical usage of wavelets has focused around the usage of discrete wavelets.

The development of statistical methods can really be grouped into three main strands of development: i) the usage of compression for non-parametric signal estimation, following work by [15,16,17], ii) the characterization of stochastic difference stationary processes, see [38,51], and iii) the inference of locally stationary wavelet processes, see [40]. Theoretical results have also been achieved in terms of the characterization of stochastic processes, see [3,4,8,34], but these results as of yet have had limited practical implications.

After initial usage of the wavelet transform for signal estimation it became obvious that despite the wavelet transform's orthogonalizing effects, relationships between coefficients must be used, to derive better representations, see [58] The development of data driven decompositions, and usage of the 'lifting algorithm', [12], followed in the second strand of developments. Further steps forward included the construction of two dimensional decompositions for image estimation, where notable contributions include the ridgelet, curvelet and bandelet transforms, see [9,33,52]. Outstanding problems remain in the statistical usage of these transforms, and investigations have mainly been limited to modeling the observed phenomena as deterministic and immersed in some form of noise.

Decomposition View of Random Processes

Wavelets are used to analyze observations whose generating mechanism can be characterized in terms of the location the observation is made at. This could be time, space or some other suitable indexing. We shall denote the location of any observation by $t \in \mathbb{R}^d$, where d denotes the dimension of the indexing, this could be $d = 1$, like a time series, or $d = 2$, like an image. If $\{x(t)\}$ is a random process then its structure can be determined from its moments. The first order structure of $x(t)$ is given by its mean, defined by

$$\mu_x(t) = \mathrm{E}\{x(t)\}\,, \quad t \in \mathbb{R}^d\,, \tag{1}$$

and the second order structure of $x(t)$ is given by its covariance, namely

$$\gamma_x(t_1, t_2) = \operatorname{cov}\{x(t_1), x(t_2)\}\,, \quad t_1,\ \ t_2 \in \mathbb{R}^d\,. \tag{2}$$

$\mathrm{E}\{\cdot\}$ denotes the expectation operator, and $\operatorname{cov}\{\cdot,\cdot\}$ the covariance operator. If the process is assumed to be a Gaussian process, i. e. if any sample of n observations from $\{x(t)\}$ is multivariate Gaussian, then $\{\mu_x(t)\}$ in combination with $\{\gamma_x(t_1, t_2)\}$ fully characterize the generation of the process $\{x(t)\}$.

Often we only take n observations from the process $\{x(t)\}$ and still want to fully characterize the generating mechanism, so that we can test hypotheses of interest regarding $\{x(t)\}$ or just estimate some parameters of interest. To be able to do this we must make some further assumptions about $x(t)$, and these could correspond to fully parametric or non-parametric assumptions on $\{\mu_x(t)\}$ or $\{\gamma_x(t_1, t_2)\}$. Typical examples might be $\mu_x(t) = a_x + b_x t$, which is a parametric specification of the mean, or $\mu_x(t)$ possessing a certain degree of smoothness. If the latter approach is taken then it is convenient to represent $\mu_x(t)$ in terms of some functions where the assumed smoothness can be easily incorporated into the modeling. Furthermore the covariance structure can be assumed to simplify if we choose a suitable representation. If $x(t)$ is represented in terms of linear combinations of known basis functions, such as wavelets, then Gaussian processes will be transformed into Gaussian processes with a different mean and covariance function. For certain classes of auto-covariance functions the analysis is much simplified by considering the transformed rather than the original Gaussian process.

Wavelets

A wavelet function, is a complex-valued square integrable function satisfying the admissibility condition, see [27]. We denote a generic wavelet by $\psi(t)$. The Continuous

Wavelet Transform (CWT) of a signal $x(t)$ with respect to a wavelet $\psi(t)$ is defined as a filtration of $x(t)$:

$$w^{(x)}(t,s) = \int_{-\infty}^{\infty} x(t')\psi_{t,s}(t')\,\mathrm{d}t' \,,$$
$$\psi_{t,s}(t') = s^{\frac{-1}{2}}\psi\left(\frac{t'-t}{s}\right) . \tag{3}$$

The choice of normalization, here $s^{-1/2}$ varies across application fields. Sometimes 1/2 is replaced by 1 if more convenient for analysis, see [54]. The former is referred to as the *energy normalization*, and the latter as the *compression normalization*. The former normalization conserves the variance of the transformed process.

If $\{x(t)\}$ is a stochastic process satisfying some regularity conditions, see [34], then $\left\{w^{(x)}(t,s)\right\}$ is a doubly indexed stochastic process with mean

$$\mathrm{E}\left\{w^{(x)}(t,s)\right\} = \int_{-\infty}^{\infty} \mu_x(t')\psi_{t,s}(t')\,\mathrm{d}t' \,,$$
$$\psi_{t,s}(t') = s^{\frac{-1}{2}}\psi\left(\frac{t'-t}{s}\right) \tag{4}$$

$$\equiv w^{(\mu_x)}(t,s) \,. \tag{5}$$

The form of this expression simplifies if $x(t)$ takes the simpler form of a time-varying oscillation, see [13], or a singularity [28,55], or is a sufficiently regular function, see [56]. Under such circumstances very few of $\left\{w^{(\mu_x)}(t,s)\right\}$ are non-zero and inferences can be made about the model from the few non-zero coefficients. In the special case of time-varying oscillations the theory of *wavelet ridge analysis* has been developed to characterize the local oscillations. Applications of such methods include mechanical vibratory signals, see [31,59], seismic signals [44,45], as well as ocean eddy time series, see [6,35].

The usage of continuous wavelets does not extend greatly beyond the aforementioned methods of signal inference. The characterization of various stochastic processes have been investigated using continuous wavelets, i. e. [3,8,25,32,34], but this has been put to little practical usage. Some interesting developments for the prediction of continuous signals using wavelets have been developed in [1,2].

The main focus of wavelet usage in statistics has instead evolved around the usage of the Discrete Wavelet Transform (DWT). The DWT is defined as a reversed convolution between a discretely sampled signal and the wavelet filter. We denote the jth level wavelet filter $\left\{h_{j,l}\right\}$, and the scaling filter $\left\{g_{j,l}\right\}$. For a complete review of the properties of these objects see [11,37,46]. These are constructed from the $j=1$ level filters $\{h_l\}$ and $\{g_l\}$, that satisfy the quadrature mirror relationship of $g_l = (-1)^{l+1}h_{L-1-l}$ where L is the length of the two filters. Given a sampled process $x_n = x(n\Delta t)$ the wavelet coefficients are defined by:

$$w_{j,l}^{(x)} = \sum_n h_{j,l}x_n \,, \quad h_{j,l} = \sum_{k=0}^{L-1} h_k g_{j-1,l-2^{j-1}k} \,, \tag{6}$$

where the wavelet filters given $\left\{g_{j-1,l}\right\}$ can now be determined iteratively. To complement the wavelet coefficients of x_n, additionally the scaling coefficients are computed, of

$$v_{j,l}^{(x)} = \sum_n g_{j,l}x_n \,, \quad g_{j,l} = \sum_{k=0}^{L-1} g_k g_{j-1,l-2^{j-1}k} \,, \tag{7}$$

this completing the iterative specification of the wavelet and scaling filters. If we collect the wavelet and scaling coefficients in a vector $\mathbf{W}^{(x)} = [w_{1,0}^{(x)}, \ldots, w_{1,N/2-1}^{(x)}, w_{2,0}^{(x)}, \ldots, v_{J/2,0}^{(x)}]$, then the DWT can be represented as:

$$\mathbf{W}^{(x)} = \mathcal{W}\mathbf{X} \,, \tag{8}$$

where $\mathcal{W}$ is composed of $\left\{h_{j,l}\right\}$ and $\left\{g_{J,l}\right\}$, see [46]. Equation (8) is not the method to compute the DWT, popularly the transform is implemented using Mallat's pyramid algorithm, see [37], but it enables the easy determination of the stochastic properties of the wavelet transform. The CWT and the DWT are not two disparate objects: if we possess a sequence of coefficients $\left\{v_{0,l}^{(x)}\right\}$ such that

$$x(t) = \sum_{n=-\infty}^{\infty} v_{0,l}^{(x)}\phi_{0,k}(t) \tag{9}$$

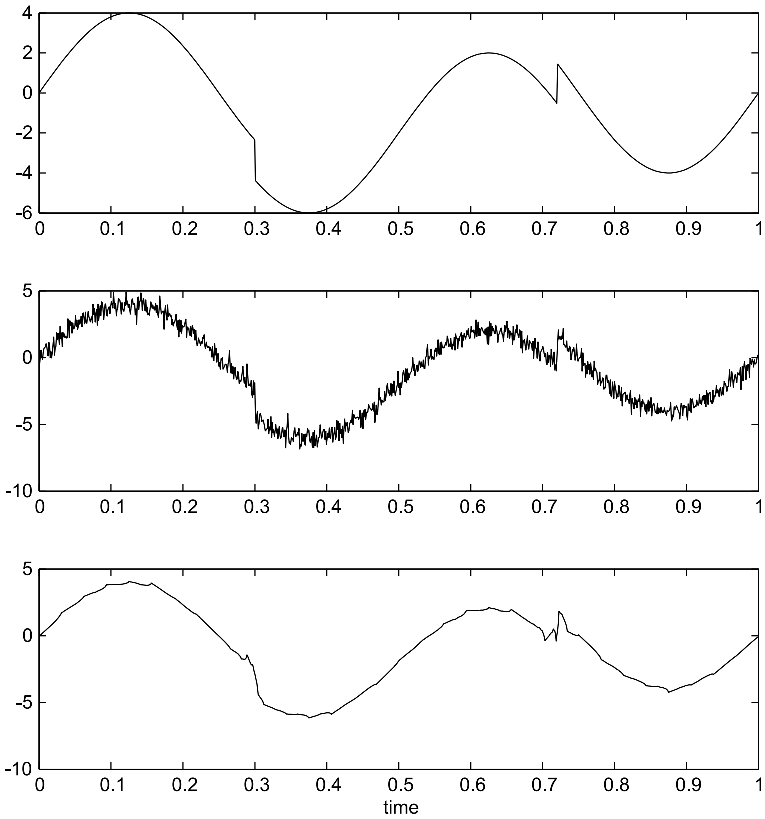
then $w_{j,l}^{(x)}$ can be calculated from $v_{0,l}^{(x)}$, as can $v_{j,l}^{(x)}$, and correspond to the CWT as well as scaling coefficients of the continuous rather than the sampled process. In practise x_n is equated to $v_{0,n}^{(x)}$, even if this is formally inappropriate.

Signal Estimation or Denoising

The most popular statistical application of wavelets is denoising, or function estimation. This proceeds from a model of the observed signal of:

$$x_n = \mu(n\Delta t) + \epsilon_n, \quad n = 0,\ldots,N-1 \,, \tag{10}$$

where Δt is referred to as the sampling period of the signal, and the sample size is N. To fully specify the generation of the observations $\{\epsilon_n\}$ must be modeled, and a popular model is to assume this sequence as zero-mean, uncorrelated (white) and Gaussian, with some constant vari-



Statistical Applications of Wavelets, Figure 1

The clean Heavisine signal (*top*), the noisy Heavisine signal (*middle*) and the denoised Heavisine signal (*bottom*). The plots show the typical features of a wavelet estimator, namely the adaptation to local smoothness, where the sinusoidal parts of the signal are smoothed, and the discontinuities, are kept without over smoothing

ance σ^2, that needs to be determined to implement estimation. To see a typical realization of such a signal, see Fig. 1. This signal, 'Heavisine' is famous as one of the four typical signals that were introduced by [16]. Heavisine has been immersed in noise, see Fig. 1. This signal is not trivial to smooth as it possesses varying degrees of smoothness across the realization, compare for example the discontinuities with the smooth portions of the signal.

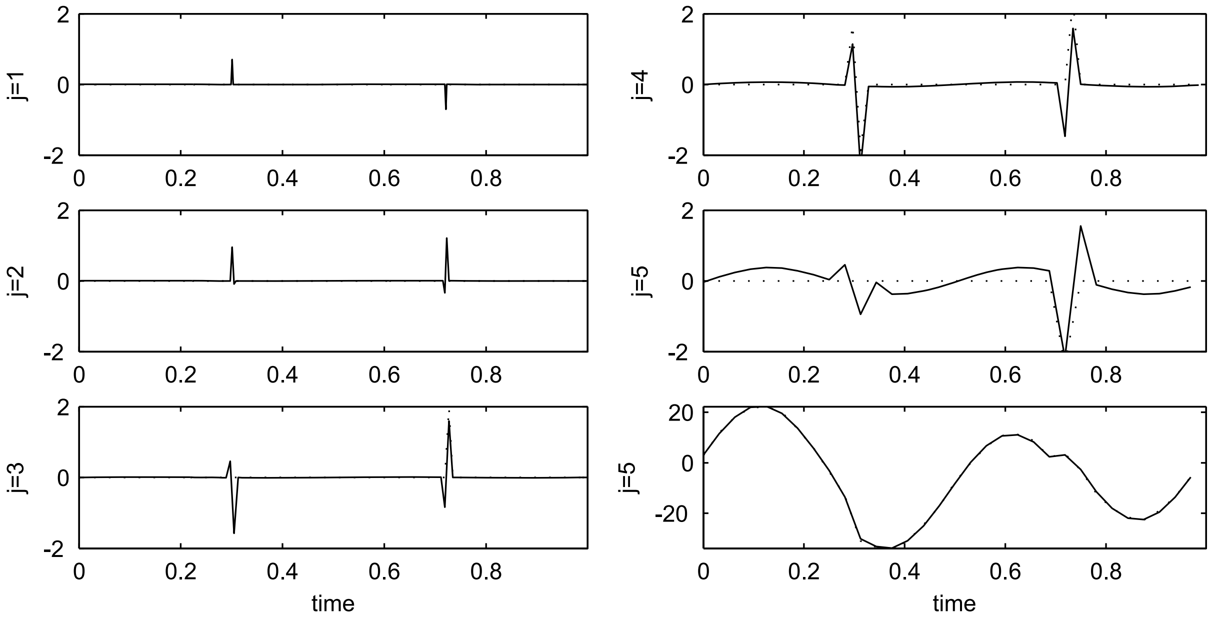
One important inference problem in this setting is to estimate $\{\mu(n\Delta t)\}$ with low expected square deviation, i. e. to chose an estimator $\widehat{\mu}(n\Delta t)$ such that

$$\mathrm{MSE}(\widehat{\mu}, \mu) = \sum_{n=0}^{N-1} \mathrm{E}\left\{\left(\mu(n\Delta t) - \widehat{\mu}(n\Delta t)\right)^2\right\} , \qquad (11)$$

is as small as possible. Common methods of estimating $\mu(n\Delta t)$ include using many different adaptive linear estimation methods, such as kernel smoothers, smoothing splines, truncated Fourier methods etc. Smoothing using the DWT was introduced by [15,16], and has enjoyed an incredible success. The simplest form of wavelet estimation is based on the compression of the DWT of $x(t)$ alone, combined with the uniformity of the white process across time.

The wavelet coefficients of $\mu(\cdot)$ are given as $\left\{w_{j,l}^{(\mu)}\right\}$, and together with the scaling coefficients, $\left\{v_{j,l}^{(\mu)}\right\}$, these are sufficient to reconstruct $\mu(n\Delta t)$ via calculating the Inverse DFT (IDFT) or:

$$\begin{pmatrix} \mu(0\Delta t) \\ \dots \\ \mu((N-1)\Delta t) \end{pmatrix} = \mathcal{W}^{-1} \begin{pmatrix} w_{1,0}^{(\mu)} \\ \dots \\ w_{1,\frac{N}{2}-1}^{(\mu)} \\ w_{2,0}^{(\mu)} \\ \dots \\ w_{2,\frac{N}{4}-1}^{(\mu)} \\ \dots \\ v_{J,0} \end{pmatrix} . \qquad (12)$$



Statistical Applications of Wavelets, Figure 2

The wavelet coefficients of the Heavisine signal, used in the original paper by Donoho and Johnstone [15]. We show the first five level wavelet coefficients, and the scaling coefficients of the signal in *solid line*. Added to the picture are the thresholded wavelet and scaling coefficients of a noisy realization in *dotted line*

Thus a good estimator of $\left\{w_{j,l}^{(\mu)}\right\}$ can easily be converted into a good estimator of $\mu(n\Delta t)$, using Eq. (12). To see the motivation behind the most commonly adopted choice of estimation for the wavelet coefficients, observe the wavelet coefficients of the Heavisine signal, that we plotted in Fig. 1, plotted in Fig. 3. Most of the wavelet coefficients are quite small. If a given signal $\mu(\cdot)$ is sufficiently regular then it can be shown that $\left\{w_{j,l}^{(\mu)}\right\}$ must decrease in magnitude. For this reason one may argue that most of this set are zero, and the transform exhibits compression or sparsity, see [56]. We are therefore trying to estimate a sequence of variables of which most are zero, and only a few are large. It is *not* optimal to equate the estimator to the sample DWT, but instead shrinkage estimation is suitable. Consider the generic setting of an estimator for w denoted by $\widehat{w}$ which is unbiased, but potentially has large variance. A shrinkage estimator of this quantity is given by $\widehat{w}^{(s)} = c\widehat{w}$ for some $0 \leq c \leq 1$. Clearly this estimator has smaller variance as compared to $\widehat{w}$ (or possibly equal variance), but introduces some bias into the estimation. The mean square error is in fact given by:

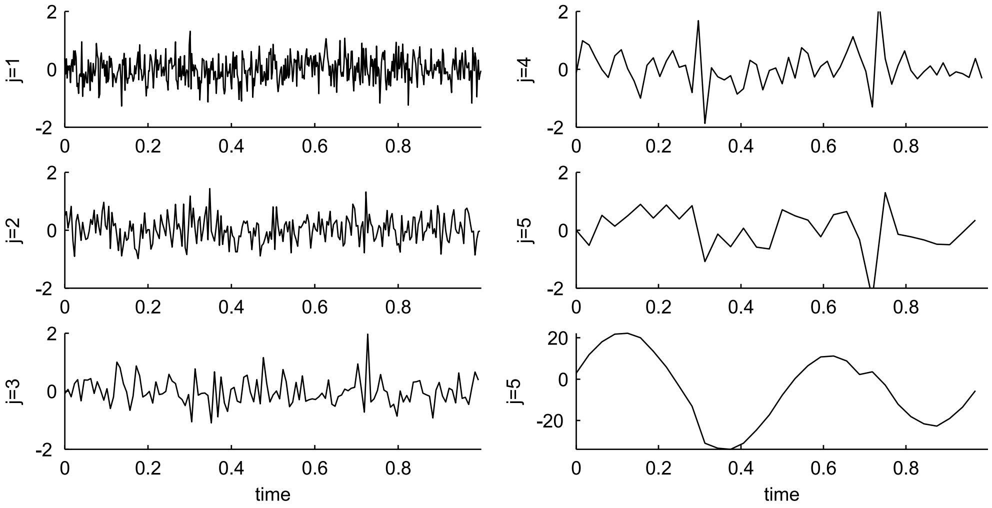
$$\mathrm{MSE}(\widehat{w}, w) = c^2\sigma^2 + (1-c)^2 w^2\,. \tag{13}$$

Thus if the quantity we are trying to estimate is small compared to the variance of the empirical estimator $\widehat{w}$ then by choosing $|c|$ smaller than one, we reduce the mean square error in estimation. Thus for $|w|$ taking a range of values it would be better to use the shrinkage estimator. This may seem non-intuitive, but corresponds to a whole set of theory derived due to [53]. If there was an oracle that could for a given shrinkage rule inform the estimator how coefficients should be shrunk, then good estimators would be obtained. Fortunately it is not necessary to have an oracle, for large sample sizes, to expect to get an estimator with equivalent performance to using an oracle. Looking at Fig. 2 and Fig. 3 we certainly see that replacing most of the empirical wavelet coefficients by zero, should greatly improve the estimation of these coefficients.

Two special shrinkage functions are commonly used, namely the *soft* and *hard* threshold functions. The soft threshold estimator modifies the empirical estimator $\widehat{w}$ by

$$\widehat{w}_{\lambda}^{(st)} = \begin{cases} \widehat{w} - \lambda & \text{if} \quad \widehat{w} \geq \lambda \\ 0 & \text{if} \quad |\widehat{w}| < \lambda \\ \widehat{w} + \lambda & \text{if} \quad \widehat{w} \leq -\lambda \end{cases} \tag{14}$$

i. e. if the observed empirical estimate is sufficiently large a set amount is removed from its magnitude, whilst if it is not sufficiently large, the estimator is set to zero. Also



Statistical Applications of Wavelets, Figure 3

The noisy wavelet coefficients of the Heavisine signal, used in the original paper by Donoho and Johnstone. The decomposition has been stopped at level five, and then the scaling coefficients have been calculated to complete the representation

popular is the hard threshold estimator, defined by:

$$\widehat{w}_{\lambda}^{(ht)} = \begin{cases} \widehat{w} & \text{if} \quad \widehat{w} \geq \lambda \\ 0 & \text{if} \quad |\widehat{w}| < \lambda \\ \widehat{w} & \text{if} \quad \widehat{w} \leq -\lambda \end{cases} . \tag{15}$$

Most wavelet coefficients of sufficiently regular functions (i. e. functions in a suitable Besov space) will be very small, or even zero. One would greatly reduce the mean square error of the estimated mean function if one estimated those by zero, rather than their observed value. Thus we define the "oracle risk" in estimation as the mean square error which arises from using the best keep or kill procedure on the empirical wavelet coefficients, denoted by R_N:

$$R_N = \sum_{n=0}^{N-1} \min\left(w_n^2, \sigma^2\right) \tag{16}$$

If instead the estimator $\widehat{w}_{\sigma\lambda_N}^{(ht)}$ for $\lambda_N = \sigma\sqrt{2\log(N)}$ (the *universal threshold*) is used, then Donoho and Johnstone showed that the mean square error of this procedure is close to the oracle risk, see [56]. Alternative methods chose a value of λ that is permitted to vary with the value of j. σ is estimated using a robust scale estimator on the first level wavelet coefficients, such as the median absolute deviation estimator.

If there is only noise present, then using the universal threshold with a hard thresholding procedure will ensure that noise only signals are estimated as zero, see [56]. [18] give a more flexible definition of a universal thresholding procedure in terms of the probability of keeping the largest noise coefficient.

The simplest wavelet estimation method, suitable for regularly spaced, uncorrelated and Gaussian data has subsequently been extended to deal with irregularly spaced data, that is contaminated with correlated noise, and is immersed in non-Gaussian perturbations. A whole class of methods arise from transforming non-Gaussian observations, using say the Anscombe transform. Of particular note is the Haar–Fisz transform, introduced by [21]. The wavelet Haar–Fisz transform is defined by:

$$f_j^k = \begin{cases} 0 & \text{if} \quad v_{jk} = 0 \\ \frac{w_{jk}}{\sqrt{v_{jk}}} & \text{if} \quad v_{jk} \neq 0 \end{cases} , \tag{17}$$

and the distribution of this variable is substantially more Gaussian than the original wavelet coefficients. Fryzlewicz and Nason have established the properties of using this transform in signal estimation in a series of papers, see [21,22,23]. The denoising of correlated data was treated by [30].

Many different extensions also exist to the basic procedure for white Gaussian random variables: usually the DWT coefficients exhibit some form of structured behavior, and either clustering across time (l) or persistence across scale (j) is observed in the magnitude of the coefficients. Various methods have been proposed to use these observed characteristics, such as treating whole blocks of coefficients together, see [7], using Bayesian methods where the dependence of the coefficients is modeled in the prior and other methods, such as wavelet footprints due to [19], and 'Analytic' denoising proposed by Olhede and Walden and later extended to two dimensions, see [43] and [42]. To observe the clustered structure of wavelet coefficients, see Fig. 2. The discontinuity in the considered signal causes wavelet coefficients to be non-zero near the discontinuities in time across all scales, and across scales at the given time location of the discontinuity. To represent the discontinuities perfectly, and avoid wavelet Gibb's effects, all of these coefficients should be kept, thus estimating a single discontinuity without artifacts. Similar statements can be made for other features than perfect discontinuities, and the above mentioned methods all address this problem in a chosen manner.

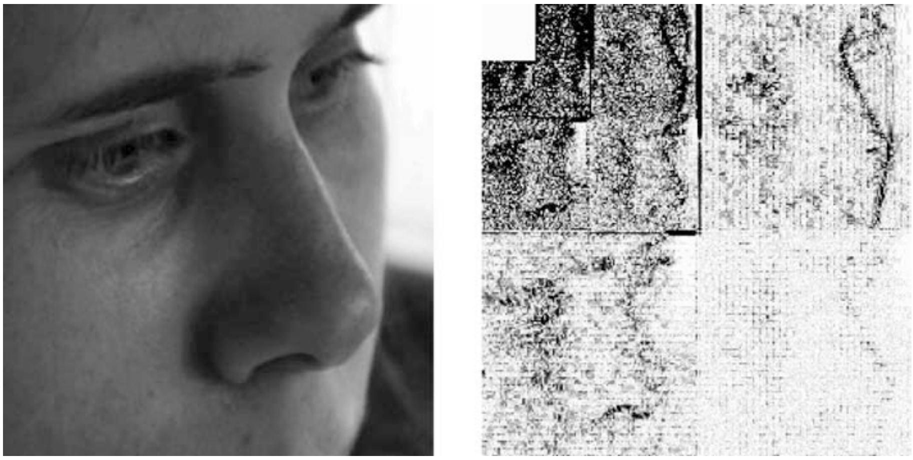
A key characteristic of observed coefficients is the sparsity level, i. e. how many of the coefficients that are approximately zero in expectation. Choosing the threshold λ should be done in light of the sparsity of the data. Optimally such sparsity is learned from the data. [29] proposed the EBAYES procedure whereby at each scale the sparsity of the data is estimated and a threshold chosen using marginal likelihood.

2-D Extensions of Wavelets

The development of statistical estimation methods for 1-D signals observed in noise is a mature field. Greater challenges lie in higher dimensional objects such as spatial, or spatiotemporal signals. In the early days when methods for extended into 2-D, the two spatial directions were treated as if the observed phenomena were two different phenomena. This caused a number of unpalatable effects in the reconstruction of the estimated images, such as edge effects and blocking, as well as using too many coefficients to reproduce simple phenomena. To see an example of this, observe Fig. 4. The woman is mainly a smooth picture, and should only correspond to a limited set of coefficients in a compressed representation. Using separable wavelets to represent the picture, because the curve of the woman's nose is misaligned with the decomposition, causes too many coefficients to be used. This lack of compression, leads to poor performance in the estimation.

The solution to these problems came by the development of non-separable decompositions, this leading to fewer non-zero coefficients in the decompositions, or by statistical processing of the separable decompositions, using the combined magnitude of the many coefficients to improve estimation. The main methods of note are ridgelets and curvelets proposed by [9,52], the dualtree complex wavelet transform, proposed by Kingsbury (see [48]), and bandelets (see [33]). The implementation of several of these transforms is slightly more complicated than the 1-D wavelet transform, but all correspond to linear operations on the observed signal. This means that the statistical properties of the transform coefficients can be determined by direct (if sometimes numerical) calculations. Again, thresholding and inverse transformation forms the basis for most estimation procedures, much in the same manner as the 1-D estimation procedures.

Open areas of research in this general setting, is the development of higher dimensional processing, and suitable transformations for discrete analysis on the sphere.



Statistical Applications of Wavelets, Figure 4

A picture of a woman (*left*), and the absolute value of the separable 2-D wavelet coefficients of the picture (*right*). Observe that the nose of the woman takes too many coefficients to record, despite the curve of the nose corresponding to a smooth curve. This illustrates the lack of compression of the 2-D separable wavelet transform

Covariance Estimation Using Wavelets

Wavelets have been used to estimate properties of zero-mean stochastic processes, characterized by their dependence structure alone. One of the fundamental aspects is noting classes of stochastic processes for which the transform provides a simplification of the description of the signal. One set of developments has defined a set of locally stationary processes using wavelet filters as the building block, so called Locally Stationary Wavelet Processes, see [41]. The estimation of the fundamental representation of this process, which is very much like a time varying spectrum, was addressed by Nason and his coauthors, and has subsequently been improved upon by Fryzlewicz and Nason, see [23]. The model has been used to classify sleep states of babies, and has been used in modeling financial times series, see [5,20]. This corresponds to the most natural modification of existing theory for locally stationary time series from local Fourier to wavelet bases.

The usage of wavelets for estimating time series that when differenced are stationary has also seen substantial investigation. By using wavelets the number of differencing steps needs often not be determined, which is a clear advantage. Likelihood theory has been extended from the Fourier domain to the wavelet domain for such time series, see [39], and various procedures for estimating the decay of long memory processes have been adapted for wavelet coefficients, see [38]. Characterization of sets of time series that are jointly stationary when differenced, has also been developed, see [49,50,51,57]. A fundamental method of analysis is to characterize the covariances of the process associated with different time-scales, this yielding a multiresolution analysis of the stochastic process, both considering the auto-covariance of a single process, and cross-covariances between several processes.

Future Directions

The usage of the wavelet transform in 1-D is a mature field. Current developments are mainly in the application and adaption of existing methods to novel applications. The development of higher dimensional transforms has also slowed down considerably. The modern thrust and drive is instead found in using several concepts which the development of analysis methods from the wavelet transform were founded upon. The key two notions for these developments are sparsity and incoherence. For many signals of interest, their wavelet coefficient representation was compressed. This means the sparsity could be used to substantially improve the estimation procedure. The area of 'compressed sensing' is based on using the compression of a signal in a known decomposition, to improve the estimation of the signal, see [14]. Surprisingly the limitations of Nyquist sampling has been circumvented, and near miraculous reconstruction performance obtained. As the data deluge that faces modern signal processing will increase, this stands as a very interesting area of future development.

Bibliography

1. Antoniadis A, Paparoditis E, Sapatinas T (2006) A functional wavelet-kernel approach for time series prediction. J Roy Stat Soc B 68:837–857
2. Antoniadis A, Sapatinas T (2003) Wavelet methods for continuous-time prediction using Hilbert-valued autoregressive processes. J Multivar Anal 87:133–158
3. Averkamp R, Houdre C (1998) Some distributional properties of the continuous wavelet transform of random processes. IEEE Trans Info Theory 44:1111–1124
4. Averkamp R, Houdre C (2000) A note on the discrete wavelet transform of second-order processes. IEEE Trans Info Theory 46:1673–1676
5. Bellegem SV, Fryzlewicz P, von Sachs R (2003) A wavelet-based model for forecasting non-stationary processes. Inst Phys Conf Ser 173:955–958
6. Buresti G, Lombardi G, Bellazzini J (2004) On the analysis of fluctuating velocity signals through methods based on the wavelet and hilbert transforms. Chaos Solitons Fractals 20:149–158
7. Cai T, Silverman BW (2001) Incorporating information on neighbouring coefficients into wavelet estimation. Sankhyā Ser B 63:127–148
8. Cambanis S, Houdre C (1995) On the continuous wavelet transform of second-order random processes. IEEE Trans Info Theory 41:628–642
9. Candès E, Donoho DL (1999) Ridgelets: a key to higher-dimensional intermittency? Philos Trans: Math Phys Eng Sci 357:2495–2509
10. Daubechies I (1988) Orthonormal bases of compactly supported wavelets. Comm Pure Appl Math 41:909–996
11. Daubechies I (1992) Ten lectures on wavelets. SIAM, Philadelphia
12. Daubechies I, Sweldens W (1998) Factoring wavelet transforms into lifting steps. J Fourier Anal Appl 4:247–269
13. Delprat N, Escudié B, Guillemain P, Kronland-Martinet R, Tchamitchian P, Torresani B (1992) Asymptotic wavelet and gabor analysis: extraction of instantaneous frequencies. IEEE Trans Inform Theory 38:644–64
14. Donoho DL (2006) Compressed sensing. IEEE Trans Inf Theory 52:1289–1306
15. Donoho DL, Johnstone IM (1994) Ideal spatial adaptation by wavelet shrinkage. Biometrika 81:425–455
16. Donoho DL, Johnstone IM (1995) Adapting to unknown smoothness via wavelet shrinkage. J Am Stat Assoc 90:1200–1224
17. Donoho DL, Johnstone IM, Kerkyacharian G, Picard D (1995) Wavelet shrinkage – asymptotia. J Roy Stat Soc B 57:301–337
18. Downie TR, Silverman BW (1998) The discrete multiple wavelet transform and threshold methods. IEEE Trans Signal Process 46:2558–2561

19. Dragotti PL, Vetterli M (2003) Wavelet footprints: theory, algorithms, and applications. IEEE Trans Signal Process 51:1306–1323
20. Fryzlewicz P, Bellegem SV, von Sachs R (2003) Forecasting nonstationary time series by wavelet process modelling. Ann Inst Stat Math 55:737–764
21. Fryzlewicz P, Nason GP (2004a) A Haar–Fisz algorithm for poisson intensity estimation. J Comp Graph Stat 13:621–638
22. Fryzlewicz P, Nason GP (2004b) A Haar–Fisz algorithm for poisson intensity estimation. J Comp Graph Stat 13:621–638
23. Fryzlewicz P, Nason GP (2006) Haar–Fisz estimation of evolutionary wavelet spectra. J Roy Stats Soc B 68:611–634
24. Grossman A, Morlet J (1984) Decomposition of Hardy functions in square integrable wavelets of constant shape. Siam J Math Anal 15:723–736
25. Guérin C-A (2000) Wavelet analysis and covariance structure of some classes of non-stationary processes. J Fourier Anal Appl 6:403–425
26. Haar A (1910) On the theory of orthogonal function systems. Math Ann 69:331–371
27. Holschneider M (1995) Wavelets: an analysis tool. Oxford University Press, Oxford
28. Jaffard S, Melot S (2005) Wavelet analysis of fractal boundaries. Commun Math Phys 258:513–159
29. Johnstone I, Silverman BW (2005) Empirical Bayes selection of wavelet thresholds. Ann Stat 33:1700–1752
30. Johnstone IM, Silverman BW (1997) Wavelet threshold estimators for data with correlated noise. J Roy Stat Soc Ser B 59:319–351
31. Kim IK, Kim YY (2005) Damage size estimation by the continuous wavelet ridge analysis of dispersive bending waves in a beam. J Sound Vib 287:707–722
32. Krim H, Pesquet J-C (1995) Multiresolution analysis of a class of nonstationary processes. IEEE Trans Inf Theory 41:1010–1020
33. le Pennec E, Mallat S (2005) Sparse geometric image representations. IEEE Trans Image Process 14:423–438
34. Li T-H, Oh H-S (2002) Wavelet spectrum and its characterization property for random processes. IEEE Trans Inf Theory 48:2922–2937
35. Lilly JM, Gascard J-C (2006) Wavelet ridge diagnosis of time-varying elliptical signals with application to an oceanic eddy. Nonlinear Process Geophys 13:467–483
36. Loève M (1945) Sur le fonctions aléatoire de second ordre. Rev Sci 83:297–303
37. Mallat S (1999) A Wavelet Tour of Signal Processing, 2nd edn. Academic Press, New York
38. McCoy EJ, Walden AT (1996) Wavelet analysis and synthesis of stationary long-memory processes. J Comp Graph Stat 5:26–56
39. Moulines E, Roueff F, Taqqu MS (2007) On the spectral density of the wavelet coefficients of long-memory time series with application to the log-regression estimation of the memory parameter. J Time Ser Anal 28:155–187
40. Nason GP, von Sachs R (1999) Wavelets in time-series analysis. Philos Trans Math Phys Eng Sci 357:2511–2526
41. Nason GP, von Sachs R, Kroisandt G (2000) Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. J Roy Stat Soc B 62:271–292
42. Olhede S (2007) Hyperanalytic denoising. IEEE Trans Image Process 16:1522–1537
43. Olhede S, Walden AT (2004a) 'Analytic' denoising. Biometrika 91:955–973
44. Olhede SC, Walden AT (2004b) The Hilbert spectrum via wavelet projections. Proc Roy Soc Lond A 460:955–975
45. Olhede SC, Walden AT (2005) Wavelet denoising for signals in quadrature. Integr Computer-Aided Eng 12:109–117
46. Percival DB, Walden AT (2000) Wavelet methods for time series analysis. Cambridge University Press, Cambridge
47. Schuster A (1898) On the investigation of hidden periodicities with application to a supposed 26-day period of meteorological phenomena. Terr Mag Atmos Elect 3:13–41
48. Selesnick IW, Baraniuk RG, Kingsbury NG (2005) The dual-tree complex wavelet transform. IEEE Signal Process Mag 22(6):123–151
49. Serroukh A, Walden AT (2000a) Wavelet scale analysis of bivariate time series i: motivation and estimation. J Nonparametr Stat 13:1–36
50. Serroukh A, Walden AT (2000b) Wavelet scale analysis of bivariate time series ii: statistical properties for linear processes. J Nonparametr Stat 13:37–56
51. Serroukh A, Walden AT, Percival DB (2000) Statistical properties and uses of the wavelet variance estimator for the scale analysis of time series. J Am Stat Assoc 95:184–196
52. Starck JL, Candès EJ, Donoho DL (2002) The curvelet transform for image denoising. IEEE Trans Image Process 11:670–84
53. Stein C (1981) Estimation of the mean of a multivariate normal distribution. Ann Stat 9:1135–1151
54. Torrence C, Compo GP (1998) A practical guide to wavelet analysis. Bull Am Meteorol Soc 79:61–78
55. Tu C-L, Hwang W-L, Ho J (2005) Analysis of singularities from modulus maxima of complex wavelets. IEEE Trans Inf Theory 51:1049–1062
56. Wasserman L (2006) All of nonparametric statistics. Springer, New York
57. Whitcher B, Guttorp P, Percival DB (2000) Wavelet analysis of covariance with application to atmospheric time series. J Geophys Res Atmos 105(14):14941–14962
58. Wolfe PJ, Godsill SJ, Ng WJ (2004) Bayesian variable selection and regularisation for time-frequency surface estimation. J Roy Stat Soc Ser B 66:575–589
59. Zhang Z, Ren Z, Huang W (2003) A novel detection method of motor broken rotor bars based on wavelet ridge. IEEE Trans Energy Convers 18:417–423

Statistical and Non-linear Physics, Introduction to

M. Cristina Marchetti
Physics Department, Syracuse University, Syracuse, USA

The field of Statistical and Nonlinear Physics combines the venerable subject of statistical mechanics with the newer area of nonlinear science to create a highly interdisciplinary and exciting area of research. It has its roots in equilibrium statistical physics, but it has evolved to encompass and emphasize nonequilibrium and dynamical phenomena. It is a field in rapid evolution, with a constantly changing focus. It overlaps naturally with many other disciplines, including fluid dynamics, computational physics, biological physics, condensed matter physics and polymer physics.

Statistical physics aims to describe the large-scale collective behavior of systems composed of a large number of interacting units or degrees of freedom. Starting with a microscopic model, one performs various types of "coarse-graining" to describe phenomena that occur on length and time scales large compared to microscopic ones, such as the size of the particles or the characteristic interaction times. At this large scale the system exhibits collective or emergent behavior that is far richer than that of the individual units. One of the hallmarks of collective behavior is phase transitions. An everyday example is the change from liquid to solid that occurs upon tuning a parameter such as temperature or pressure. Another familiar example arises in the study of the properties of magnetic materials and is highlighted in the article ▶ Complex Systems and Emergent Phenomena, which is a good starting point for the reader of this section.

When going from the deterministic Hamiltonian description of a system of many interacting units to the large scale coarse-grained description of the same system in terms of a continuum or hydrodynamic theory, an element of randomness is naturally introduced into the problem as one looses the invariance under time reversal that was present in the Hamiltonian description. At the same time randomness on the other hand is an intrinsic property of the dynamics of individual nonlinear systems that are known to often exhibit sensitive dependence to initial conditions and chaotic behavior. There is in fact a deep connection between the randomness of chaotic systems and the irreversible transport properties of extended systems (see ▶ Chaotic Dynamics in Nonequilibrium Statistical Mechanics). This connection highlights the unity of the two areas of statistical and nonlinear science.

Rather than presenting an exhaustive review of the field of statistical and nonlinear physics, the articles in this section focus on nonequilibrium phenomena that are the subject of current research. Even then, only a fraction of the problems and physical systems that are studied using the tools and ideas of statistical and nonlinear physics are described here. Many of the others can be found in articles throughout the rest of the Encyclopedia.

The topics highlighted in this section of the Encyclopedia may at a first glance seem disparate, but are unified by the ideas that are used to study them. Principal among those are the notions of scaling and universality. The concept of universality, which has been around for some time, has its roots in the study of phase transitions and critical phenomena. Near a phase transition, the system is universal in that its behavior at large scales does not depend on the microscopic physics. The systems and physical problems described in this section of the Encyclopedia may be different in the details, but generally involve the onset of collective or emergent behavior with "universal" properties.

The ideas of scaling universality have had a crucial impact in the study of long-standing problems in nonlinear physics, such as pattern formation and turbulence (see ▶ Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence). The same tools and ideas have more recently been applied to the study of the complex dynamics of neuronal systems (see ▶ Neuronal Dynamics), lasers (see ▶ Noise and Stability in Modelocked Soliton Lasers), and even the physics of economics (see ▶ Econophysics, Statistical Mechanics Approach to). Rich chaotic behavior has also been discovered in quantum systems (see ▶ Quantum Chaos).

Recently, a new field has emerged that uses many of the ideas of statistical mechanics to explore problems in "soft" condensed matter physics. Soft materials are literally materials that yield easily to mechanical deformations, They encompass an extremely wide range of systems, from complex fluids such as colloidal suspensions, liquid crystals and polymers (see ▶ Polymer Physics) to soft elastic gels (see ▶ Anisotropic Networks, Elastomers and Gels), and granular matter in both its fluidized (see ▶ Granular Flows) and its jammed (see ▶ Jamming of Granular Matter) states. The jamming or structural arrest of granular matter when the density reaches a critical value bears striking similarities with the glass transition. Jammed granular matter and glassy systems share many of the intriguing dynamical properties discussed in ▶ Glasses and Aging, A Statistical Mechanics Perspective on. Jamming also provides an example of the nonequilibrium phase transitions that are ubiquitous in the systems studied by modern statistical physics. Nonequilibrium phase transitions do occur in "pure" systems, i. e., ones without any extrinsic disorder, as a result of the dynamical constraints arising from interactions, but are even more common in systems with quenched disorder, induced for instance by impurities and materials defects (see ▶ Disordered Elastic Media for a general perspective on such systems). Notable examples include the depinning transition that occurs in driven extended systems such as vortex lattices in type-II superconductors under the action of a uniform external driving force (see ▶ Collective Transport and Depinning) and the slip-stick motion that characterizes the physics of friction and tectonic motion (see ▶ Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of). Living matter, such as the cytoskeleton of eukaryotic cell or even the whole cell, has also been described as a novel form of nonequilibrium soft matter (see ▶ Cytoskeleton and Cell Motility).

The complexity of the systems that are studied by statistical physicists has naturally led to the development of sophisticated numerical methods. Principal among those is the technique of Monte Carlo simulations (see ▶ Monte Carlo Simulations in Statistical Physics). More recently physicists have also begun to adapt ideas and algorithms from computer science (see ▶ Optimization Problems and Algorithms from Computer Science).

Novel experimental techniques have recently allowed us to begin to study the properties of soft and living matter on length and time scales that require a qualitative new way of thinking and a modification of the familiar tools from statistical physics (see ▶ Fluctuation Theorems, Brownian Motors and Thermodynamics of Small Systems). These range from optical techniques for study of the properties of individual biomolecules (see ▶ Protein Mechanics at the Single-Molecule Level) to microfluidics devices for the study of the rheological and mechanical properties of complex materials and cells on novel time and length scales (see ▶ Microfluidics).

Yet another class of problems that has caught the attention of statistical physicists lately is the study of networks (see ▶ Networks: Structure and Dynamics). Examples of networks are as disparate as the world-wide web, human societies, languages and their structure, and power grids. An especially important class of networks occurs in biology. It consists of the cell signaling networks that control many of the functions of living cells (see ▶ Cell Biology: Networks, Regulation and Pathways).

Finally, complex emergent behavior is also ubiquitous in quantum matter, where the interactions and fluctuations are controlled by quantum mechanics, as described in the article ▶ Ultracold Atomic Gases: Novel States of Matter.

Statistics with Imprecise Data

María Ángeles Gil[1], Olgierd Hryniewicz[2]
[1] Faculty of Sciences, University of Oviedo, Oviedo, Spain
[2] Systems Research Institute, Warsaw, Poland

Article Outline

Glossary

Fuzzy estimators Estimators of "parameters" of probability distributions, or other characteristics, of random variables/fuzzy random variables (such as e. g. the expected value/the fuzzy expected value) when statistical data are imprecise and are described by means of fuzzy sets.

Fuzzy random variable Random element whose observed values are described by fuzzy sets.

Fuzzy set Generalization of a classical notion of a set. In contrast to the case of a classical set, each element x of a fuzzy set may belong to it to a degree described by the so-called membership function $\mu(x)$. Thus, the fuzzy set may be defined as a set of ordered pairs $(x, \mu(x))$, where x belongs to a set $\mathbb{X}$ called the universe of discourse or referential. Alternatively, the fuzzy set can be identified with its membership function (in the same way that a classical set can be identified with its indicator function).

Fuzzy statistical tests Statistical tests used for the verification of hypotheses about the values of "parameters" of probability distributions, or other characteristics, of random variables/fuzzy random variables when statistical data are imprecise and are described by fuzzy sets.

Fuzzy statistics Generalization of traditional statistics that allows to handle imprecise data described in terms of fuzzy sets.

Imprecise data Data that cannot be described by either real numbers or vectors with real-valued components.

Random sets Random elements whose observed values are sets (e. g. intervals, subsets of a plane).

Definition of the Subject

Traditional statistics deals usually with precisely defined data. The source of these data may be of different nature. Usually the data are collected from observations of random experiments, i. e. experiments whose performance leads to an outcome which cannot be predicted in advance with one hundred percent sureness. When the knowledge about the nature of these random events is available we can use the methods of mathematical statistics and draw conclusions about the source of available data. Uncertainty being intrinsic to random outcomes/events is properly described by using the formalism of the mathematical theory of probability, and generally it is attributed to *future* observations of these outcomes/events.

Traditionally, statistical experimental data are described by real numbers or by vectors whose components are real numbers. These numbers are either observed directly as results of measurements (e. g. height and weight of a person) or correspond to observed counts of certain categories representing labeled events (e. g. a gender of that person). However, in real life applications the results of measurements are never precise. The precision of every measurement is limited by the precision of a measuring device. In the majority of practical applications this lack of precision may be neglected, and for statistical analysis of available data we can use traditional methods of statistics. If the measurement error cannot be neglected statistical analysis of interval data is recommended by specialists in metrology. However, when statistical data are presented by human beings we need more sophisticated methods for the description of their lack of precision. Therefore, there is a practical need to generalize traditional statistical methods in order to make them applicable for handling imprecise data, e. g. the data described by statements expressed by using a plain language (the so-called "linguistic data").

Introduction

There is a common agreement that uncertainty characterized by randomness shall be described by considering the theory of probability. Thus, mathematical statistics is a proper tool for dealing with data generated by random experiments and described by precisely defined numbers. However, there also exist other types of uncertainty which are related to vagueness, imprecision, existence of only partial information about experimental outcomes/events of interest, etc. In contrast to randomness, uncertainties of such types are attributed rather to *current* perceptions/observations. It has to be noted that the mathematical modeling of all these types of uncertainty which are different from simple randomness is still the subject of controversies. A good overview of the related problems can be found in the paper by P. Walley [36]. In this article we confine ourselves to the case when randomness is observed together with vagueness understood as the lack of precision.

Specialists in measurement theory recognize different types of uncertainty. For instance, in the ISO/IEC Guide [18] it is recommended to distinguish between uncertainty related to pure randomness and uncertainty of other nature, such as a lack of precision. However, the lack of precision of statistical data is usually omitted in the statistical analysis of measurements. Consider, for example, the analysis of results of measurements coming from a digital meter. The results of measurements coming from such a meter are always rounded in order to have their representation by a certain number of digits. When we observe a displayed result of a measurement we never know what is the actual value of the measured quantity. What is more important, and often overlooked by statisticians, we may not increase our knowledge about that value by averaging the results of repeated measurements, as it is frequently recommended by statisticians. Consider, for example, a series of measurements showing exactly the same result. Having such results we are in principle not able to distinguish between two fundamentally different cases when the measured quantity is constant and its value belongs to the range of rounding or when it is varying from measurement to measurement with values belonging to that range.

The situation described in the previous paragraph is relatively simple as the range of possible actual values corresponding to the observed result of a measurement is usually precisely defined. There exist, however, situations when either this range is not precisely known or values in the range are not equally compatible with the available data/information/events. We face such cases when results of measurements or classifications are evaluated by humans (e. g., by evaluating indications of an analog meter or classifying individuals in accordance with their height, and so on). The result of a measurement is even more imprecise when it has been obtained without the usage of any meter (e. g., when we visually evaluate the distance between two points); in such situations statistical data may consist of imprecise statements like "around 5 meters", "more or less between 5 and 10 seconds", etc. The classification of a person in accordance with his/her height leads often to imprecise assessments, like "very short", "rather tall", etc. A similar situation occurs when we deal with retrospective data recalled by human beings; for example, in reliability analysis of field lifetime data we may face situations when failure times are reported imprecisely using statements like 'the failure occurred about one month ago', etc. In all these cases statistical data consist of *imprecise perceptions* of actual real values.

In the previous paragraphs we considered situations when actual values of measured quantities exist, but they are imprecisely perceived. There exist, however, situations when we have to analyze statistical data that represent imprecisely defined concepts. Take for example the color of human hair described using categories such as "blond" or "dark blond"; it is obvious that the border between these two categories is vague; one can try to establish a precise border between these two categories in terms of the results of precise measurements of the spectrum of reflected light, but such attempts seem to be practically senseless. We face a similar situation in classifying a client of a bank office in accordance with his/her degree of aversion to investment which leads to imprecise assessments, like "very low", "moderate", etc. In both these cases we could try to collect precise statistical data, e. g., by either asking a respondent to a questionnaire to indicate exactly one choice or coding it numerically. However, it seems to be more prudent, natural and informative to expect imprecise answers to questions pertained to vague notions. If we do so, we may face imprecise statistical data for further analysis. In all these cases statistical data consist of *imprecise actual values* themselves.

There also exists another source of imprecision while dealing with statistical data. For example, in reliability lifetime tests we perform tests in more severe "over-stress" conditions, and then we try to recalculate test results to conditions which are considered "normal". For this recalculation we can utilize some partial knowledge about possible values of stress-dependent recalculation coefficients. In such a case we have originally precise lifetime data, but after recalculation these data become imprecise.

In all considered cases the lack of precision can be appropriately described by using fuzzy sets introduced by Lotfi A Zadeh. In the second section of the article we recall some basic definitions related to fuzzy sets. We will present the fuzzy sets methodology as a useful tool for the description of imprecise data. When fuzzy lack of precision is mixed with randomness, either in the sense that available fuzzy data are supposed to come from the perception of real- or vectorial-valued data generated by a random mechanism, or in the sense of they being directly generated by a random mechanism, a convenient tool to use is that of the notion of a fuzzy random variable. This notion is introduced in the third section of this article. The interpretation of the fuzzy random variable depends upon the type of observed fuzzy data and events. In the paper we distinguish between the two types of data which have been described in this section. We start with the description of statistical methods which are useful for the analysis of fuzzy perceptions of existing precise values. Then, we present statistical methods which are useful in the analysis of intrinsically fuzzy-valued data.

Mathematical Modeling of Imprecise Data

There exist competitive methods for the description of vagueness. For example, some statisticians claim that the theory of subjective probability is sufficient for the description of all types of uncertainty. However, many other researchers have shown examples of situations when the application of the classical theory of probability is not sufficient for modeling these situations. Therefore, other formalisms have been proposed for the description of vagueness/imprecision. One of those formalisms, namely the theory of *fuzzy sets* proposed by Lotfi A. Zadeh [37], has been slowly but widely accepted as a good methodology for the description of imprecise data, both from practical and theoretical (see, e. g., the paper by Terán [33]) points of view.

The basic concept of the theory of fuzzy sets is the universe of discourse or referential $\mathbb{X}$ which may be understood as the set of all possible (feasible) elements that are relevant for the description of a certain concept (quantity). Mathematically speaking a ***fuzzy subset*** A of a set $\mathbb{X} \neq \emptyset$ (or a fuzzy set, for short) is a map $A : \mathbb{X} \to [0, 1]$, where $A(x)$ can be interpreted as the degree of compatibility of x with the ill-defined property characterizing A, or degree of truth of the assertion "x is A", or degree of membership of x to A. Equivalent, but more intuitive for some purposes, a fuzzy set A of $\mathbb{X}$ can be defined as a set of ordered pairs $\{(x, \mu_A)\}$, where $x \in \mathbb{X}$ and $\mu_A : \mathbb{X} \to [0, 1]$ is the so-called *membership function* of A. In other words, a fuzzy set can be identified with its membership function, in the same way that a classical set can be identified with its indicator function. In what follows, we will consider indistinctly A or μ_A to denote and refer to a fuzzy subset.

Unfortunately, there is no one generally accepted methodology for the construction of membership functions. The majority of researchers assume that the membership function $\mu_A(x)$ is a purely subjective function provided by a person who describes his/her perception of a certain phenomenon or quantity. Some authors provide practical methods for the construction of the membership function when it is interpreted in terms of the theory of possibility as the possibility distribution (see [8] for more information). Some other authors, e.g Bandemer and Näther [1] or Viertl [35], present methods which may be used for the construction of membership functions in a more objective way from measurements of physical quantities. Anyway, our purpose is not entering a discussion here about this point.

In the analysis of imprecise data we are usually interested in the description of interesting phenomena by numbers. For this purpose we can use the concept of a fuzzy number defined as follows (see [5]):

Definition 1 The fuzzy subset A of the space of real numbers $\mathbb{R}$, with the membership function $\mu_A : \mathbb{R} \to [0, 1]$, is a **fuzzy number** if and only if

(a) A is normal, i. e. there exists at least an element $x_0 \in \mathbb{R}$ such that $\mu_A(x_0) = 1$;
(b) A is fuzzy convex, i. e., $\mu_A(\lambda x_1 + (1-\lambda)x_2) \geq \mu_A(x_1) \wedge \mu_A(x_2)$, for all $x_1, x_2 \in \mathbb{R}$, and $\lambda \in [0,1]$;
(c) μ_A is upper semicontinuous;
(d) the support set, $\operatorname{supp} A = \{x \in \mathbb{R} : \mu_A(x) > 0\}$, is bounded (that is, its closure $\operatorname{cl}(\operatorname{supp} A)$ is compact).

It is easily seen that if A is a fuzzy number then its membership function can be expressed as follows:

$$\mu_A(x) = \begin{cases} 0 & \text{for } x < a_1 \\ r_{l_A}(x) & \text{for } a_1 \leq x < a_2 \\ 1 & \text{for } a_2 \leq x \leq a_3 \\ r_{u_A}(x) & \text{for } a_3 < x \leq a_4 \\ 0 & \text{for } x > a_4 \,, \end{cases} \tag{1}$$

where $a_1, a_2, a_3, a_4 \in \mathbb{R}$, $a_1 \leq a_2 \leq a_3 \leq a_4$, $r_{l_A} : [a_1, a_2] \to [0,1]$ is a nondecreasing upper semicontinuous function, and $r_{u_A} : [a_3, a_4] \to [0,1]$ is a nonincreasing upper semicontinuous function. Functions r_{l_A} and r_{u_A} are called sometimes the left and the right "arms" (or "sides") of the fuzzy number, respectively.

A useful notion for dealing with a fuzzy number is the so-called α-level set, also known as the *α-cut*. For $\alpha \in (0,1]$ the *α-level set* of the fuzzy number A is the ordinary (non-fuzzy) set defined as

$$A_\alpha = \{x \in \mathbb{R} : \mu_A(x) \geq \alpha\} \tag{2}$$

and the 0-level set is usually intended to be given by $A_0 = \operatorname{cl}(\operatorname{supp} A)$. The family $\{A_\alpha : \alpha \in [0,1]\}$ is a set representation of the fuzzy number A. According to the resolution identity proposed by Zadeh, we can represent the membership function as:

$$\mu_A(x) = \sup\{\alpha \cdot \mathbf{1}_{A_\alpha}(x) : \alpha \in (0,1]\}\,, \tag{3}$$

where $\mathbf{1}_{A_\alpha}(x)$ denotes the indicator function of A_α.

On the basis of the notion of α-level, a fuzzy number A can be viewed as a fuzzy subset of $\mathbb{R}$ such that its α-level sets are nonempty compact and convex sets of $\mathbb{R}$, that is, nonempty compact intervals. Hence, for each $\alpha \in [0,1]$ we have that $A_\alpha = [A_L(\alpha), A_U(\alpha)]$, where

$$\begin{aligned} A_L(\alpha) &= \inf\{x \in \mathbb{R} : \mu_A(x) \geq \alpha\}\,, \\ A_U(\alpha) &= \sup\{x \in \mathbb{R} : \mu_A(x) \geq \alpha\}\,. \end{aligned} \tag{4}$$

If the sides of the fuzzy number A are strictly monotonic functions, then from Eq. (1) one can see easily that $A_L(\alpha)$ and $A_U(\alpha)$ are inverse functions of r_{l_A} and r_{u_A}, respectively.

In statistical analysis of random data we use functions (and, in particular, operations) of the observed random samples. These functions, called statistics, can be also defined for fuzzy random data. Their membership functions can be derived by using Zadeh's *extension principle*. This principle has the following formulation [38]:

Let $\mathbb{X}$ be a Cartesian product of universes $\mathbb{X} = \mathbb{X}_1 \times \ldots \times \mathbb{X}_r$, and $A_1, \ldots, A_r$ be r fuzzy subsets of $\mathbb{X}_1, \ldots, \mathbb{X}_r$, respectively. Let f be a mapping from $\mathbb{X} = \mathbb{X}_1 \times \ldots \times \mathbb{X}_r$ to a universe $\mathbb{Y}$ such that $y = f(x_1, \ldots, x_r)$. The extension principle allows us to induce from the r fuzzy sets A_i a fuzzy set B on $\mathbb{Y}$ through f, $B = f(A_1, \ldots, A_r)$ such that

$$\mu_B(y) = \begin{cases} \sup_{x_1,\ldots,x_r | y = f(x_1,\ldots,x_r)} \\ \quad \min\{\mu_{A_1}(x_1), \ldots, \mu_{A_r}(x_r)\} & \text{if } f^{-1}(y) \neq \emptyset \\ 0 & \text{if } f^{-1}(y) = \emptyset \end{cases} \tag{5}$$

In case of A_i, $i = 1, \ldots, n$ being fuzzy numbers, and f being either an injective or a continuous function, then for each $\alpha \in [0,1]$ the α-level of $B = f(A_1, \ldots, A_r)$ can be shown to be equal to $B_\alpha = \big(f(A_1, \ldots, A_r)\big)_\alpha$ with

$$\begin{aligned} &\big(f(A_1, \ldots, A_r)\big)_L(\alpha) \\ &\quad = \min_{(x_1,\ldots,x_r) \in (A_1)_\alpha \times \ldots \times (A_r)_\alpha} f(x_1, \ldots, x_r)\,, \\ &\big(f(A_1, \ldots, A_r)\big)_U(\alpha) \\ &\quad = \max_{(x_1,\ldots,x_r) \in (A_1)_\alpha \times \ldots \times (A_r)_\alpha} f(x_1, \ldots, x_r)\,. \end{aligned} \tag{6}$$

Thus, the application of the extension principle for the calculation of the membership function of $y = f(x_1, \ldots, x_r)$ is equivalent to the application of the interval arithmetics on α-level sets of the arguments of this function (see [29] as a basis for the proof). For instance, if A and B are fuzzy numbers, then,

$$\begin{aligned} (A+B)_L(\alpha) &= A_L(\alpha) + B_L(\alpha), \\ (A+B)_U(\alpha) &= A_U(\alpha) + B_U(\alpha)\,, \\ (\lambda \cdot A)_L(\alpha) &= \begin{cases} \lambda \cdot A_L(\alpha) & \text{if } \lambda \geq 0 \\ \lambda \cdot A_U(\alpha) & \text{if } \lambda < 0 \end{cases} \\ (\lambda \cdot A)_U(\alpha) &= \begin{cases} \lambda \cdot A_U(\alpha) & \text{if } \lambda \geq 0 \\ \lambda \cdot A_L(\alpha) & \text{if } \lambda < 0 \end{cases} \end{aligned}$$

for any $\lambda \in \mathbb{R}$, and $\alpha \in [0,1]$.

Fuzzy Random Variables

Uncertainty, understood as randomness, is well described in Probability Theory. The concept of a random variable, which is basic in this theory, is well-known and its definition does not need to be recalled in this article.

However, when we observe random experimental data which are imprecise, a useful tool to model either the imprecise perception of values coming from real-valued random variables or the random mechanisms generating directly these imprecisa data is the one associated with the so-called concept of fuzzy random variables. Actually, we can consider two different approaches to the concept of fuzzy random variable; the motivation for these approaches and the situations they apply to are different, but the formalization of the second notion and the associated statistical methodology can be applied to the first one.

Historically, the first widely accepted definition of the fuzzy random variable was proposed by Kwakernaak [21,22]. Kruse [19] proposed an interpretation of this notion, and according to this interpretation a fuzzy random variable $\mathcal{Z}$ may be considered as a fuzzy perception of an unknown true real-valued random variable Z_0 associated with a random experiment, and referred to as 'the original' of $\mathcal{Z}$. Below, we recall the version of this definition elaborated by Kruse and Meyer [20].

Definition 2 (Kruse and Meyer [20]) Let $(\Omega, \mathcal{A}, P)$ be a probability space, where Ω is the set of all possible outcomes of a random experiment, $\mathcal{A}$ is a σ-field of subsets of Ω (the set of all possible events of interest), and P is a probability measure associated with $(\Omega, \mathcal{A})$.

A mapping $\mathcal{X} : \Omega \to \mathcal{F}_c(\mathbb{R})$, where $\mathcal{F}_c(\mathbb{R})$ is the space of all fuzzy numbers, is called a **fuzzy random variable** if it satisfies the following properties:

i) $\{\mathcal{X}_\alpha(\omega) : \alpha \in [0,1]\}$, where $\mathcal{X}_\alpha(\omega) = (\mathcal{X}(\omega))_\alpha$ is a set representation of $\mathcal{X}(\omega)$ for all $\omega \in \Omega$;
ii) for each $\alpha \in [0,1]$ both $\mathcal{X}_\alpha^L : \Omega \to \mathbb{R}$ and $\mathcal{X}_\alpha^U : \Omega \to \mathbb{R}$, with $\mathcal{X}_\alpha^L(\omega) = \inf \mathcal{X}_\alpha(\omega)$ and $\mathcal{X}_\alpha^U(\omega) = \sup \mathcal{X}_\alpha(\omega)$, are usual real-valued random variables associated with $(\Omega, \mathcal{A}, P)$.

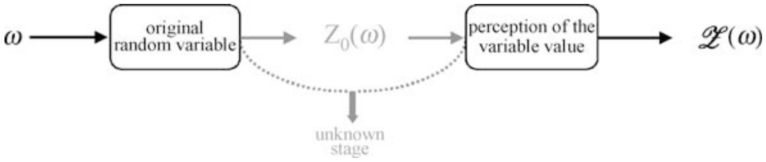
Values of a fuzzy random variable in Definition 2 have been conceived to model fuzzy perceptions of existing real-valued values (the values of the original, see Fig. 1). For instance, when we qualify the price of a given item in a specific store, we can perceive/label it as being 'cheap', but there is an existing price (although assumed to be unknown for the person receiving and processing data information).

Puri and Ralescu [31] introduced the concept of the also called fuzzy random variable as a generalization of the concept of random set or set-valued random element (and hence, as a generalization also of the concept of random variable). According to this definition a fuzzy random variable $\mathcal{Z}$ may be considered as a random element associating with each experimental outcome a value which is intrinsically fuzzy. Below, we recall Puri and Ralescu's definition

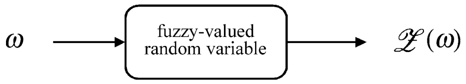
Definition 3 (Puri and Ralescu [31]) Given a probability space $(\Omega, \mathcal{A}, P)$, a mapping $\mathcal{X} : \Omega \to \mathcal{F}_c(\mathbb{R})$ is said to be a *fuzzy random variable* (also referred to as *random fuzzy set*) if for each $\alpha \in [0,1]$ the set-valued mapping $X_\alpha : \Omega \to \mathcal{K}_c(\mathbb{R})$, where $\mathcal{K}_c(\mathbb{R})$ is the class of the nonempty compact intervals and $X_\alpha(\omega) = (X(\omega))_\alpha$ for all $\omega \in \Omega$, is a compact convex random sets (that is, a Borel-measurable mapping with respect to the Borel σ-field generated by the topology associated with the Haussdorf metric on $\mathcal{K}_c(\mathbb{R})$).

Remark 1 The notion of fuzzy random variable has been in fact introduced in a more general way by considering as the codomain the space of fuzzy sets of the p-dimensional Euclidean space, or even more general Banach spaces, whose α-levels are nonempty compact subsets of this space. In case one constrains to $p = 1$ and fuzzy sets being convex, then one gets the last definition.

Remark 2 Although motivation to introduce fuzzy random variables was different in the approaches by Kwakernaak/Kruse and Meyer and by Puri and Ralescu, one can prove that the notion in Definition 3 implies the one in Definition 2. As a consequence, probabilistic ideas and results for the notion in Definition 3 (or for the more general one in Remark 1) apply to the notion in Definition 2, and the same happens for statistical developments. However, many probabilistic conclusions, and most of the statistical procedures for Definition 2 are based on the assumption



Statistics with Imprecise Data, Figure 1
Fuzzy random variables in Kwakernaak/Kruse and Meyer's sense



Statistics with Imprecise Data, Figure 2
Fuzzy random variables in Puri and Ralescu's sense

of having an unknown but existing original, and considering Zadeh's extension principle, so that these conclusions and procedures are not usually applicable to deal with data coming from fuzzy random variables in Definition 3.

Remark 3 The concept of fuzzy random variable in Definition 3 can be alternatively formalized as a Borel-measurable mapping with respect to the Borel σ-field generated by the topology associated with some metrics on the space $\mathcal{F}_c(\mathbb{R})$, among them an operational one we will later refer to. Borel-measurability allows us to guarantee that notions like those of the induced distribution by a fuzzy random variable, independence of fuzzy random variables, identically distributed fuzzy random variables, and so on, can be immediately formalized in the probabilistic setting.

Values of a fuzzy random variable in Definition 3 have been conceived to model existing fuzzy values (see Fig. 2). For instance, when we classify a client of a bank in accordance with the degree of aversion to investment as having a 'rather high' degree, there is no underlying real-valued degree, but the classification itself is essentially imprecise.

Statistical Analysis of Fuzzy Data Corresponding to Fuzzy Perceptions of Existing Real–Valued Data

Fuzzy Estimation

When imprecise statistical data correspond to fuzzy perceptions of unobserved/unknown precise (i. e. crisp) statistical data we can treat them as observed values of fuzzy random variables in the sense of Kwakernaak/Kruse and Meyer. In such a case we can analyze imprecise data in terms of probability distributions of their unobserved originals in a similar way as precise statistical data are analyzed using methods of traditional mathematical statistics. The only difference stems from the fact that having imprecise input information in the form of fuzzy data we cannot precisely evaluate the characteristics of the underlying probability distribution. Therefore, instead of finding precise values of the estimators of the 'parameters' describing the underlying probability distribution (the one of the original), it seems more coherent finding their imprecise fuzzy perceptions.

Assume that we observe a fuzzy random sample $\mathcal{X}_1,\dots,\mathcal{X}_n$ which is viewed as a fuzzy perception of an unobserved random sample $X_1,\dots,X_n$. Let $F(x;\theta)$ be the cumulative probability function of the original random variable X characterized by a crisp parameter $\theta\in\Theta$. Suppose now that an estimator of θ which is given by a statistic $\hat{\theta}=\phi(X_1,\dots,X_n)$ is considered. By using Zadeh's extension principle we can consider as a *fuzzy estimator* of θ based on $\mathcal{X}_1,\dots,\mathcal{X}_n$ the one associating with each fuzzy sample information $(\tilde{x}_1,\dots,\tilde{x}_n)$ the *fuzzy estimate* $\phi(\tilde{x}_1,\dots,\tilde{x}_n)$ given by

$$\mu_{\phi(\tilde{x}_1,\dots,\tilde{x}_n)}(t) = \begin{cases} \sup_{(x_1,\dots,x_n)\,|\,t=\phi(x_1,\dots,x_n)} \min\{\mu_{\tilde{x}_1}(x_1),\dots,\mu_{\tilde{x}_n}(x_n)\} \\ \qquad \text{if } t\in\mathrm{Im}(\phi(X_1,\dots,X_n) \\ 0 \quad \text{otherwise} \end{cases} \tag{7}$$

Alternatively, $\phi(\widetilde{x}_1,\dots,\widetilde{x}_n)$ can be viewed as a fuzzy estimate of the induced fuzzy parameter $\vartheta=\theta(\mathcal{X})$ with $\theta(\mathcal{X})(t)=\sup_{X\in\mathcal{E}(\Omega,\mathcal{A},P)\,|\,\theta(X)=t}\inf_{\omega\in\Omega}\mu_{\mathcal{X}(\omega)}X(\omega)$ and $\mathcal{E}(\Omega,\mathcal{A},P)$ being the class of all possible originals of $\mathcal{X}$. In many practical cases, when $\phi(x_1,\dots,x_n)$ has a simple form, the calculation of (7) is straightforward. For example, when $\phi(x_1,\dots,x_n)=$ sample mean $=(x_1+\ldots+x_n)/n$ the α-levels of the estimate of its expected value θ are expressed as

$$(\phi(\tilde{x}_1,\dots,\tilde{x}_n))_\alpha=\left[\frac{1}{n}\sum_{i=1}^{n}(\tilde{x}_i)_L(\alpha),\ \frac{1}{n}\sum_{i=1}^{n}(\tilde{x}_i)_U(\alpha)\right] \tag{8}$$

Analogously, whenever the estimator of θ is given by a continuous function or an injective function, the α-levels becomes also rather simple. In other cases, the limits of the α-levels should often be found by solving nonlinear mathematical programming problems defined by (7).

A similar approach may be applied when we try to construct fuzzy versions of the confidence intervals of the unknown parameter θ. Let us assume that we are able to find confidence intervals of θ using precise (crisp) statistical data. For example, let $[\underline{\pi}_l,+\infty)$, where $\underline{\pi}_l=\underline{\pi}_l(X_1,\dots,X_n;\delta)$, be the one-sided confidence interval for the parameter θ on the confidence level $1-\delta$. Kruse and Meyer [20] have shown that when we replace $\underline{\pi}_l$ with the lower limits of the α-cuts of its fuzzy version we obtain a proper confidence interval for the fuzzy perception of θ. These lower limits can be found for different values of $\alpha\in(0,1]$ from the formula [15]

$$\begin{aligned}\underline{\Pi}^L_\alpha &= \underline{\Pi}^L_\alpha(\mathcal{X}_1,\dots,\mathcal{X}_n;\delta)\\ &=\inf\big\{t\in\mathbb{R}:\forall i\in\{1,\dots n\}\ \exists x_i\in\mathcal{X}_{i,\alpha}\\ &\quad\text{such that }\underline{\pi}_l(x_1,\dots,x_n;\delta)\le t\big\}\end{aligned} \tag{9}$$

where $X_{i,\alpha}, i = 1, \ldots, n$ are respective α-cuts of the observed fuzzy data.

In a similar way we can define a fuzzy equivalent of the one-sided confidence interval $(-\infty, \overline{\pi}_u]$:

$$\begin{aligned}\overline{\Pi}_\alpha^U &= \overline{\Pi}_\alpha^U(\mathcal{X}_1, \ldots, \mathcal{X}_n; \delta) \\ &= \sup\big\{t \in \mathbb{R} : \forall i \in \{1, \ldots n\}\, \exists z_i \in X_{i,\alpha} \\ &\quad \text{such that } \overline{\pi}_u(z_1, \ldots, z_n; \delta) \geq t\big\} \qquad (10)\end{aligned}$$

where $\overline{\pi}_u(z_1, \ldots, z_n; \delta) = \underline{\pi}_l(z_1, \ldots, z_n; 1-\delta)$. Moreover, exactly the same approach can be applied when we look for two-sided confidence intervals.

Summing up this section we can say that when imprecise observations may be treated as fuzzy perceptions of precise but unobserved realizations of ordinary random variables the problem of point and interval estimation of the unknown parameters of the underlying probability distribution can be reduced to finding fuzzy versions of the formulae known from traditional mathematical statistics.

Fuzzy Statistical Tests

Testing statistical hypotheses is the second main branch of mathematical statistics. Tests of statistical hypotheses have to be applied if we want to make decisions based on the analysis of random data. When our decisions depend on the values of the parameters of probability distributions that describe observed statistical data we use parametric statistical methods. In such a case we test statistical hypotheses about the values of the parameters of probability distributions utilizing a well known equivalence between the set of values of the considered probability distribution parameter for which the null hypothesis is accepted and a certain confidence interval for this parameter. Kruse and Meyer [20] have shown that the same equivalence exists in the case of statistical tests with fuzzy data.

Let $\mathcal{X}_1, \ldots, \mathcal{X}_n$ denote a fuzzy sample, i. e. a fuzzy perception of the usual random sample $X_1, \ldots, X_n$, from the population with the distribution P_θ. Let δ be a given number from the interval $(0, 1)$. Grzegorzewski ([15]) proposed the following definition of the *fuzzy test* for vague data:

Definition 4 A function $\varphi : \left(\mathcal{F}_c(\mathbb{R})\right)^n \to \mathcal{F}(\{0,1\})$ is called a fuzzy test for the hypothesis H, at the significance level δ, if

$$\sup_{\alpha\in[0,1]} P\{\omega \in \Omega : \varphi_\alpha(\mathcal{X}_1(\omega), \ldots, \mathcal{X}_n(\omega)) \subseteq \{1\}\,|H\} \leq \delta, \qquad (11)$$

where φ_α is the α-level set (α-cut) of $\varphi(\mathcal{X}_1, \ldots, \mathcal{X}_n)$.

The fuzzy test defined above can be regarded as a family of classical tests $\{\varphi_\alpha : \alpha \in (0,1]\}$ for which the significance level is given as the upper bound of type I error for the whole family $\{\varphi_\alpha : \alpha \in (0,1]\}$.

In order to give an example of a fuzzy statistical test let us consider a following simple null hypothesis: $H : \theta = \theta_0$, against the composite two-sided alternative: $K : \theta \neq \theta_0$. Suppose we know a two-sided symmetrical confidence interval $[\pi_1, \pi_2]$ for θ, on a confidence level $1-\delta$, where $\pi_1 = \pi_1(X_1, \ldots, X_n; \delta/2)$ and $\pi_2 = \pi_2(X_1, \ldots, X_n; \delta/2)$ are the limits of the ordinary two-sided confidence interval. The fuzzy equivalent of this confidence interval can be calculated using the α-cuts $\Pi_\alpha = [\Pi_\alpha^L, \Pi_\alpha^U]$ for all $\alpha \in (0,1]$, where the limits of these α-cuts can be computed from Eq. (9)-(10) by replacing δ with $\delta/2$. The fuzzy two-sided statistical test for $H : \theta = \theta_0$, against $K : \theta \neq \theta_0$, on the significance level δ, has been defined by Grzegorzewski [15] as a function $\varphi : (\mathcal{F}_c(\mathbb{R}))^n \to \mathcal{F}(\{0,1\})$ with following α-cuts

$$\varphi_\alpha(X_1, \ldots, X_n) = \begin{cases} \{0\} & \text{if } \theta_0 \in (\Pi_\alpha \setminus (\neg\Pi)_\alpha)\,, \\ \{1\} & \text{if } \theta_0 \in ((\neg\Pi)_\alpha \setminus \Pi_\alpha)\,, \\ \{0,1\} & \text{if } \theta_0 \in (\Pi_\alpha \cap (\neg\Pi)_\alpha)\,, \\ \emptyset & \text{if } \theta_0 \notin (\Pi_\alpha \cup (\neg\Pi)_\alpha)\,, \end{cases} \qquad (12)$$

Similarly we may obtain fuzzy tests for one-sided hypotheses using the one-to-one correspondence between the acceptance regions of the tests designated for testing these hypotheses on the significance level δ and one-sided confidence intervals for the parameter θ on the confidence level $1-\delta$.

Grzegorzewski [15] has shown that the membership function of the fuzzy test for the hypothesis H against K is given by

$$\begin{aligned}\mu_\varphi(t) &= \mu_\Pi(\theta_0)I_{\{0\}}(t) + \mu_{\neg\Pi}(\theta_0)I_{\{1\}}(t) \\ &= \mu_\Pi(\theta_0)I_{\{0\}}(t) + (1-\mu_\Pi(\theta_0))I_{\{1\}}(t)\,,\ t \in \{0,1\}, \qquad (13)\end{aligned}$$

where Π is a fuzzy acceptance region depending on the considered hypotheses. Thus, the fuzzy fuzzy test defined by Eq. (12), contrary to the classical crisp test, does not lead to the binary decision – to accept or to reject the null hypothesis – but to a fuzzy decision. One may get $\mu_\varphi(0) = 1, \mu_\varphi(1) = 0$ which indicates that we should accept H, or $\mu_\varphi(0) = 0, \mu_\varphi(1) = 1$ which means the rejection of H. However, one may also get $\mu_\varphi(0) = \mu_0, \mu_\varphi(1) = 1 - \mu_0$, where $\mu_0 \in (0,1)$, which can be interpreted as a degree of conviction that we should accept (μ_0) or reject ($1-\mu_0$) the hypothesis H. Thus, in situation when μ_0 is neither 0 nor 1, a user must decide using other criteria whether to reject or to accept the considered hypothesis. There exist several approaches that are suitable for solving this problem. One of these approaches which is formulated in the language of the possibility theory has been proposed by Hryniewicz [16] who used the results of Dubois et al. [9] who proposed to use statistical confidence intervals of parameters of probability distributions for the construction of possibility distributions of these parameters. According to their approach, the family of two-sided confidence intervals

$$\begin{aligned}&[\pi_L(x_1,\ldots,x_n;1-\delta/2)\,,\\&\qquad\pi_U(x_1,\ldots,x_n;1-\delta/2)]\,,\qquad\delta\in(0,1)\end{aligned}\tag{14}$$

forms the *possibility distribution* $\tilde{\vartheta}$ of the estimated value of the unknown parameter ϑ. In a similar way it is possible to construct one-sided possibility distributions based on one-sided nested confidence intervals. Hryniewicz [16] proposed to compare this possibility distribution with a hypothetical value of the tested parameter. For this purpose he proposed to use the necessity of strict dominance measure introduced by Dubois and Prade [7] for measuring the necessity of the strict dominance relation $\tilde{A}\succ\tilde{B}$, where $\tilde{A}$ and $\tilde{B}$ are fuzzy sets. This measure, called the *Necessity of Strict Dominance* index (*NSD*), is defined as

$$\begin{aligned}NSD&=\mathrm{Ness}\left(\tilde{A}\succ\tilde{B}\right)\\&=1-\sup_{x,y;x\le y}\min\left\{\mu_A(x)\,,\mu_B(y)\right\}\,.\end{aligned}\tag{15}$$

Hryniewicz [16] has shown that in the classical case of precise statistical data and precisely defined statistical hypotheses the value of the *NSD* index is equal to the *p*-value of the test.

In case of fuzzy data the confidence intervals used for the construction of the possibility distribution of the estimated parameter θ can be replaced by their fuzzy equivalents presented in the previous sections of this article. In his paper Hryniewicz [16] assumes that the value of the significance level of the corresponding statistical test δ is equal to the possibility degree α that defines the respective α-cut of the possibility distribution of $\tilde{\theta}$. He also assumes that in the possibilistic analysis of statistical tests on the significance level δ we should take into account only those possible values of the fuzzy sample whose possibility is not smaller than δ. Thus, the α-cuts of the membership function $\mu_F(\theta)$ denoted by $\left[\mu_{F,L}^{(\alpha)},\mu_{F,U}^{(\alpha)}\right]$ are equivalent to the α-cuts of the respective fuzzy confidence intervals on a confidence level $1-\alpha$. Having the possibility distribution of the test statistic we can use Eq. (15) for the calculation of the degree on necessity that the considered statistical hypothesis has to be accepted. When we set a critical value for this characteristic we arrive at unequivocal (crisp) decisions. It is worthy to note that this approach has been generalized in [16] to the case of testing imprecisely defined hypothesis using fuzzy statistical data.

In the previous paragraphs we have presented fuzzy statistical tests when the class the underlying probability distribution belongs to is known. Verification of this assumption when the available statistical data are imprecise may be very difficult indeed. Therefore, it would be advisable to use fuzzy equivalents of non-parametric (distribution-free) statistical methods. Unfortunately, there exist only few papers devoted to such fuzzy tests. The most interesting result has been obtained by Denœux et al. [4] who proposed a general methodology for the construction of fuzzy tests based on rank statistics.

Statistical analysis of fuzzy random data can be also done in the Bayesian framework. First results presenting the Bayesian decision analysis for imprecise data were given in papers by Casals et al. [3] and Gil [11]. Other approaches have been proposed by such authors as Viertl [34], Frühwirth-Schnatter [10], and Taheri and Behboodian [32]. Comprehensive Bayesian model comprising fuzzy data, fuzzy hypotheses, and fuzzy utility function has been proposed in the paper by Hryniewicz [17].

Statistical Analysis of Existing Fuzzy-Valued Data

When imprecise statistical data correspond to intrinsic fuzzy-valued data we can treat them as observed values of fuzzy random variables in the sense of Puri and Ralescu. In the literature these data are often treated as categorical/ordinal/interval-valued ones. It should be emphasized that the model given by fuzzy random variables allows us to describe and handle these data in a more expressive scale and way (in contrast to just ranking or stating simply the interval support of the values). Thus, many statistical developments for real-valued data are based on distances/deviations between values rather than on the diversity of these values. The use of the fuzzy scale allows to consider metrics with a meaning similar to that for the real-valued case (i. e., distinguishing not only the ranks of variable values w.r.t. a certain criterion, but a physical distance between them).

The distance we will consider here is the one stated by Bertoluzza et al. [2], so that if $A,B\in\mathcal{F}_c(\mathbb{R})$

$$D_W^{\varphi}(A,B)=\sqrt{\int_{[0,1]}\left[\int_{[0,1]}\left[\mathrm{f}_A(\alpha,\lambda)-\mathrm{f}_B(\alpha,\lambda)\right]^2\,\mathrm{d}W(\lambda)\right]\mathrm{d}\varphi(\alpha)}\tag{16}$$

with $\mathrm{f}_A(\alpha, \lambda) = \lambda \cdot A_U(\alpha) + (1-\lambda) \cdot A_L(\alpha)$, where

- W and φ are normalized weighted measures on $[0,1]$ formalized as probability measures on $([0,1], \mathcal{B}_{[0,1]})$,
- W is associated with a non-degenerate distribution,
- φ is associated with a strictly increasing distribution function on $[0,1]$.

Remark It should be remarked on D_W^φ that

- W and φ have no stochastic meaning.
- To consider W is equivalents to consider a measure weighting points 0, 1 and a certain $t_0(W) \in (0,1)$; in case W is symmetric, then $t_0(W) = .5$.
- For each α, the choice of W allows us to weight
  - the effect of the distance between the "widths" of the α- levels (i. e., effect of the "shape" difference),
  - in comparison with the effect of the distance between their $t_0(W)$-points (i. e., effect of the "location" difference).
- The choice of φ allows us to weight the influence of each level (i. e., degree of "imprecision", "consensus", "subjectivity",...).
- D_W^φ is a versatile and operational in statistical developments with fuzzy numbers, and it behaves especially well when we consider least-squares approaches.

Since the concept of fuzzy random variable in Puri and Ralescu's sense has been properly stated in a probabilistic context as a random element (i. e., as a Borel-measurable function), concepts like independent and identically distributed fuzzy random variables make immediate sense. Furthermore, all the main ideas, aims and concepts, and several developments can be immediately considered to deal with fuzzy data when coming from fuzzy-valued Borel-measurable mappings. In this respect, notions like either unbiasedness or consistency of a "point" (fuzzy- or real- valued) estimator of a (fuzzy- or real- valued) 'parameter' associated with the distribution of the fuzzy random variable, or the p-value and power of a test concerning such a 'parameter', make the same sense as in the classical case.

Several developments have been made in connection with both estimation and testing of fuzzy- and real-valued parameters associated with the distribution of a fuzzy random variable in Puri and Ralescu's sense. We are now just recalling a few results concerning the "point" fuzzy estimation and testing of the population mean of a fuzzy random variable, which is formalized as follows:

Definition 5 (Puri & Ralescu, [31]) Given a probability space $(\Omega, \mathcal{A}, P)$ and an associated fuzzy random variable $\mathcal{X} : \Omega \to \mathcal{F}_c(\mathbb{R})$ such that $\max\{|\inf \mathcal{X}_0|, |\sup \mathcal{X}_0|\}$ is integrable, then, the **fuzzy expected value** (or **fuzzy mean**) of $\mathcal{X}$ is the fuzzy number $\tilde{E}(\mathcal{X}|P) \in \mathcal{F}_c(\mathbb{R})$ such that for all $\alpha \in [0,1]$

$$\begin{aligned}
&\left(\tilde{E}(\mathcal{X}|P)\right)_\alpha \\
&= \text{Aumann integral of } \mathcal{X}_\alpha \\
&= \big\{E(X|P) \big| X : \Omega \to \mathbb{R}, \\
&\qquad\qquad X \in L^1(\Omega, \mathcal{A}, P),\ X \in \mathcal{X}_\alpha \text{ a.s. } [P]\big\} \\
&= \left[E(\inf \mathcal{X}_\alpha|P), E(\sup \mathcal{X}_\alpha|P)\right] .
\end{aligned}$$

Remark 5 The fuzzy mean satisfies that

- If $\mathcal{X}(\Omega) = \{\tilde{x}_1, \ldots, \tilde{x}_m, \ldots\} \subset \mathcal{F}_c(\mathbb{R})$, then, it is coherent with fuzzy arithmetic, that is,

$$\begin{aligned}
\tilde{E}(\mathcal{X}|P) &= P(\{\omega \in \Omega \,|\, \mathcal{X}(\omega) = \tilde{x}_1\}) \cdot \tilde{x}_1 + \ldots \\
&\quad + P(\{\omega \in \Omega \,|\, \mathcal{X}(\omega) = \tilde{x}_m\}) \cdot \tilde{x}_m + \ldots
\end{aligned}$$

- Strong Laws of Large Numbers are satisfied for different metrics (like D_W^φ, and stronger ones), which also corroborates the suitability of the defined fuzzy mean as the stochastic limit of the sample ones.
- $\tilde{E}(\mathcal{X}|P)$ is the "Fréchet expectation" of $\mathcal{X}$ w.r.t. D_W^φ, i. e., for all $A \in \mathcal{F}_c(\mathbb{R})$:

$$E\left(\left[D_W^\varphi(\mathcal{X}, \tilde{E}(\mathcal{X}|P))\right]^2 \big| P\right) \le E\left(\left[D_W^\varphi(\mathcal{X}, A)\right]^2 \big| P\right) .$$

The interpretation of the fuzzy mean becomes more clear if we notice that for interval-valued random variables (i. e., random intervals) the expected value is given in a form of an interval with the lower limit equal to the expected value of the lower limits of observed random variables, and the upper limit equal to the expected value of the upper limits of observed random variables. Therefore, the expected value of the fuzzy random variable can be given as a nested set of such intervals represented by respective α-cuts, as it is formally described in the definition.

The definition of other characteristics describing fuzzy random variables may be much more complicated. For example, the definition of the variance requires the introduction of the "Fréchet expectation" (defined above), as it was proposed by Körner and Näther [24].

Fuzzy Estimation

Assume that we observe a fuzzy simple random sample $\mathcal{X}_1, \ldots, \mathcal{X}_n$ which is viewed now as an n-tuple of n independent fuzzy random variables which are identically distributed as $\mathcal{X}$. The associated ***fuzzy sample mean*** is the statistic given by

$$\overline{\mathcal{X}}_n = \frac{1}{n} \cdot [\mathcal{X}_1 + \ldots + \mathcal{X}n] \ .$$

Then, one can prove that (see Lubiano ([25])

Theorem 1 *The fuzzy sample mean satisfies that*

i) $\overline{\mathcal{X}}_n[\cdot]$ *in an "unbiased fuzzy-valued estimator" of* $\tilde{E}(\mathcal{X}|P)$ *(in the sense of the fuzzy expected value defined by Puri & Ralescu). For most of the metrics we can consider, it is also a 'strongly consistent' fuzzy-valued estimator of* $\tilde{E}(\mathcal{X}|P)$*.*

ii) One can quantify the mean squared-type error in the fuzzy estimation by considering the real-valued expected value $E\left(\left[D_W^{\varphi}(\overline{\mathcal{X}}_n, \tilde{E}(\mathcal{X}|P))\right]^2\right)$*.*

Actually, Lubiano et al. [26,27] proposed several developments in connection with the estimation and testing about the D_W^{φ}-mean squared error associated with the estimation of the population fuzzy means by means of the sample one.

Statistical Tests

One of the problems which has received a deep attention in connection with testing from fuzzy random variables in Puri and Ralescu's sense is that concerning two-sided tests about the mean of a fuzzy random variable (one-sample case). In order to define this problem one can decide which metrics could be used for measuring the distance between the hypothetical and observed fuzzy values of considered characteristics of the fuzzy random variables. In the case of the metrics defined by Eq. (16), the problem can be formalized as follows:

Given n independent observations from $\mathcal{X}$, $\mathcal{X}_1, \ldots, \mathcal{X}_n$, we wish to test the null hypothesis $H_0\colon \tilde{E}(\mathcal{X}|P) = A \in \mathcal{F}_c(\mathbb{R})$, which can be equivalently expressed as $H_0\colon D_W^{\varphi}\left(\tilde{E}(\mathcal{X}|P), A\right) = 0$.

Despite simple formulation of the testing problem the construction of statistical tests for the verification of hypotheses related to fuzzy mean values is not that simple. For example, an exact test has been developed in Montenegro et al. [28] for so called "normal" fuzzy random variables (in Puri and Ralescu's sense, [30]). Although the method is exact and easy-to-apply, the assumption of $\mathcal{X}$ being fuzzy "normal" ($\mathcal{X} = V + \mathcal{N}(0,1)$, with $V \in \mathcal{F}_c(\mathbb{R})$) is quite restrictive and often unrealistic, as it means that *all* observed fuzzy values are described by the same membership function shape with maybe different location.

In a more general case asymptotic tests based on Large Sample Theory have been developed for the same purpose. However, they are usually hardly applicable (except for some simple special cases) in practice. For example, the asymptotic distribution of the statistic proposed by Körner [23] involves unknown parameters such as correlations between random variables describing membership functions of observed fuzzy variables. In the paper by Montenegro et al. [28] devoted to the problem of testing the equality of two fuzzy mean values, it is assumed that $\mathcal{X}$ takes on a finite number of values, and large sample sizes would be required anyway in order to estimate unknown parameters of the model.

Simulation studies have been considered to analyze the extent and applicability of the asymptotic test by Montenegro et al. [28]. These studies have confirmed that in estimating the unknown parameters (the eigenvalues of a certain correlation matrix) entails a substantial loss of precision w.r.t. the nominal significance level. Based on this empirical conclusion, the use of D_W^{φ} and the Generalized Bootstrapped CLT (Giné and Zinn, [13]) allow us to consider bootstrap techniques in this context.

The bootstrap technique consists in taking random samples (i. e. re-sampling) from the original one and calculating the value of the considered statistic. By repeating this procedure many times we obtain the sample distribution of the test statistic which may be thus used for testing purposes. In case of fuzzy random data we get the following method proposed by González-Rodríguez et al. [14]:

Theorem 2 *Given a fuzzy random variable* $\mathcal{X}\colon \Omega \to \mathcal{F}_c(\mathbb{R})$ *associated with the probability space* $(\Omega, \mathcal{A}, P)$ *and such that*

- $\max\left\{(\inf \mathcal{X}_0)^2, (\sup \mathcal{X}_0)^2\right\}$ *is integrable ,*
- $\mathcal{X}_1, \ldots, \mathcal{X}_n$ *are i.i.d. as* $\mathcal{X}$ *,*
- $\mathcal{X}_1^*, \ldots, \mathcal{X}_n^*$ *is a bootstrap sample from* $\mathcal{X}_1, \ldots, \mathcal{X}_n$ *,*

then, to test H_0 *at the nominal significance level* $\alpha \in [0,1]$*,* H_0 *should be rejected whenever*

$$\frac{\left[D_W^{\varphi}\left(\overline{\mathcal{X}_n}, U\right)\right]^2}{\widehat{S}_n^2} > z_\alpha \ ,$$

where z_α *is the* $100(1-\alpha)$ *fractile of the bootstrap distribution of*

$$T_n = \left[D_W^{\varphi}\left(\overline{\mathcal{X}_n^*}, \overline{\mathcal{X}_n}\right)\right]^2 \Big/ \widehat{S}_n^{*\,2}$$

with

$$\overline{\mathcal{X}_n^*} = \sum_{i=1}^{n} \mathcal{X}_i^*/n, \quad \widehat{S}_n^{*\,2} = \sum_{i=1}^{n} \left[D_W^{\varphi}\left(\mathcal{X}_i^*, \overline{\mathcal{X}_n^*}\right)\right]^2/(n-1) \, .$$

Other bootstrap tests about means of fuzzy random variables which have been recently developed are the following ones:

- One-sided hypotheses tests in the one-sample case.
- Tests for the equality of means of two FRVs
  - for two independent samples
  - for two linked samples.
- Tests for the equality of means of J FRVs (ANOVA):
  - for J independent samples.

From comparative extensive simulation studies that have been performed recently at the University of Oviedo and the European Center for Soft Computing we can draw the following practical conclusions:

- For small/medium samples, the bootstrap method performs and behaves much better than the asymptotic one.
- For large sample sizes (over 300), the improvement is not that remarkable, but the bootstrap approach still provides the best approximation to the nominal significance level.

Taking into account that fuzzy statistical test for testing hypotheses about other characteristics of fuzzy random variables are even more complicated than the procedures designed for testing the hypotheses about mean values we can claim that the bootstrap technique is probably the most promising for dealing with random fuzzy observations.

Future Directions

Statistical analysis of imprecise data is still a developing area of science. Future directions of its development are tightly connected with the development of methods that may be used for the description of uncertainties of different types. It has been pointed out by P. Walley (see, e. g. [36] for a good overview) that traditional probability is not sufficient for good description of different types of uncertainty. Different mathematical models, such as e. g. Dempster–Schafer belief functions, possibility distributions, lower and upper probabilities, lower and upper previsions, and many others, have been proposed for this purpose. However, for the most general models describing uncertainty appropriate statistical methods have not been proposed yet. Therefore, statistical methods for handling very general imprecise data have to be developed in the future.

Another, but much more specific, future direction for the development of statistical analysis of imprecise data is related to the analysis of intrinsically fuzzy data. In contrast to the situation when fuzzy observations may be considered as fuzzy perceptions of real-valued observations, many notions known from traditional statistics are still waiting for their widely accepted definitions, and statistical methods of analysis.

The most challenging future direction is related to Zadeh's paradigm of "computing with words". First of all, we need operational methods for the representation of linguistic concepts which could be useful is statistical analysis of imprecisely reported (with words!) statistical data. Moreover, we also need methods for convincing presentation of the results of computations to users who have only limited knowledge of mathematics and statistics.

Bibliography

Primary Literature

1. Bandemer H, Näther W (1992) Fuzzy Data Analysis. Kluwer, Dordrecht
2. Bertoluzza C, Corral N, Salas A (1995) On a new class of distances between fuzzy numbers, Mathware and Soft Computing, vol 2. Departament of Computer Science and Artificial Intellingence of the University of Granada and Secció de Matemàtiques i Informàtica of the Universitat Politècnica de Catalunya, Granada, Barcelona, pp 71–84. http://docto-si.ugr.es/Mathware/ENG/mathware.html
3. Casals R, Gil MA, Gil P (1986) The fuzzy decision problem. In: An approach to the problem of testing statistical hypotheses with fuzzy information. Eu J Oper Res 27:371–382
4. Denœux T, Masson M-H, Hébert PA (2005) Nonparametric rank-based statistics and significance tests for fuzzy data. Fuzzy Sets Syst 153:1–28
5. Dubois D, Prade H (1978) Operations on Fuzzy Numbers. Int J Syst Sci 9:613–626
6. Dubois D, Prade H (1980) Fuzzy Sets and Systems. In: Theory and Applications. Academic Press, New York
7. Dubois D, Prade H (1983) Ranking fuzzy numbers in the setting of possibility theory. Inf Sci 30:184–244
8. Dubois D, Prade H (1988) Possibility Theory. Plenum Press, New York
9. Dubois D, Foulloy L, Mauris G, Prade H (2002) Probability-possibility transformations, triangular fuzzy-sets and probabilistic inequalities, Proc of the Ninth International Conference IPMU, Annecy, pp 1077–1083
10. Frühwirth-Schatter S (1993) Fuzzy Bayesian inference. Fuzzy Sets Syst 60:41–58
11. Gil MÁ, Kacprzyk J, Fedrizzi M (eds) (1988) Probabilistic-Possibilistic approach to some Statistical Problems with Fuzzy Experimental Observations. In: Combining Fuzzy Imprecision with Probabilistic Uncertainty in Decision Making. Springer, Berlin, pp 286–306
12. Gil MÁ, López-Díaz M, Ralescu DA (2006) Overview on the development of fuzzy random variables. Fuzzy Sets Syst 157:2546–2557
13. Giné E, Zinn J (1990) Bootstrapping general empirical measures. Ann Prob 18:851–869

14. González-Rodríguez G, Montenegro M, Colubi A, Gil MA (2006) Bootstrap techniques and fuzzy random variables. Synergy in hypothesis testing with fuzzy data. Fuzzy Sets Syst 157:2608–2613
15. Grzegorzewski P (2000) Testing statistical hypotheses with vague data. Fuzzy Sets Syst 112:501–510
16. Hryniewicz O (2006) Possibilistic decisions and fuzzy statistical tests. Fuzzy Sets Syst 157:2665–2673
17. Hryniewicz O, Grzegorzewski P, Gil MÁ (eds) (2002) Possibilistic Approach to the Bayes Statistical Decisions. In: Soft Methods in Probability, Statistics and Data Analysis. Physica, Heidelberg, pp 207–218
18. ISO/IEC (1995) Guide to the expression of uncertainty in measurement (GUM). ISO/IEC, Geneva
19. Kruse R (1982) The strong law of large numbers for fuzzy random variables. Inf Sci 28:233–241
20. Kruse R, Meyer KD (1987) Statistics with Vague Data. D. Riedel Publishing Company, Dordrecht
21. Kwakernaak H (1978) Fuzzy Random Variables. Definitions and Theorems. Inf Sci 15:1–15
22. Kwakernaak H (1979) Fuzzy Random Variables. Algorithms and Examples for the Discrete Case. Inf Sci 17:253–278
23. Körner R (2000) An asymptotic α-test for the expectation of random fuzzy variables. J Stat Plann Inference 83:331–346
24. Körner R, Näther W (2002) On the variance of random fuzzy variables. In: Bertoluzza C, Gil MÁ, Ralescu DA (eds) Statistical Modeling, Analysis and Management of Fuzzy Data. Physica, Heidelberg, pp 25–42
25. Lubiano MA (1999) Medidas de variación de elementos aleatorios. Ph D Thesis, University of Oviedo
26. Lubiano MA, Gil MA (1999) Estimating the expected value of fuzzy random variables in random samplings from finite populations. Stat Pap 40:277–295
27. Lubiano MA, Gil MA, López-Díaz M, López-García MT (2000) The $\overrightarrow{\lambda}$-mean squared dispersion associated with a fuzzy random variable. Fuzzy Sets Syst 111:307–317
28. Montenegro M, Colubi A, Casals MR, Gil MA (2004) Asymptotic and Bootstrap techniques for testing the expected value of a fuzzy random variable. Metrika 59:31–49
29. Nguyen HT (1978) A note on the extension principle for fuzzy sets. J Math Anal Appl 64:369–380
30. Puri ML, Ralescu DA (1985) The concept of normality for fuzzy random variables. Ann Prob 13:1373–1379
31. Puri ML, Ralescu DA (1986) Fuzzy Random Variables. J Math Anal Appl 114:409–422
32. Taheri SM, Behboodian J (2001) A Bayesian approach to fuzzy hypotheses testing. Fuzzy Sets Syst 123:39–48
33. Terán P (2007) Probabilistic foundations for measurement modelling with fuzzy random variables. Fuzzy Sets Syst 158:973–986
34. Viertl R (ed) (1987) Is it necessary to develop a fuzzy Bayesian inference. In: Probability and Bayesian Statistics. Plenum, New York, pp 471–475
35. Viertl R (1996) Statistical Methods for Non-Precise Data. CRC Press, Boca-Raton
36. Walley P (1996) Measures of uncertainty in expert systems. Artif Intell 114:1–58
37. Zadeh LA (1956) Fuzzy sets. Inf Control 8:338–353
38. Zadeh LA (1975) The concept of a linguistic variable and its application to approximate reasoning. Inf Sci 8(1):199–249; 8(2):301–353; 9(3):43–80

Books and Reviews

Bertoluzza C, Gil MÁ, Ralescu DA (eds) (2002) Statistical Modeling, Analysis and Management of Fuzzy Data. Physica, Heidelberg

Dubois D, Lubiano MA, Prade H, Gil MÁ, Grzegorzewski P, Hryniewicz O (eds) (2008) Soft methods for handling variability and imprecision. Springer, Berlin

Gil MÁ, López-Díaz M, Ralescu DA (2006) Overview on the development of fuzzy random variables. Fuzzy Sets Syst 157:2546–2557

Grzegorzewski P, Hryniewicz O, Gil MÁ (eds) (2002) Soft methods in probability. In: Statistics and data analysis. Physica, Heidelberg

Kruse R, Gebhardt J, Gil MÁ (1999) Fuzzy Statistics. In: Webster JG (ed) Encyclopedia of Electrical and Electronics Engineering. Wiley, New York

Lawry J, Miranda E, Bugarin A, Li S, Gil MÁ, Grzegorzewski P, Hryniewicz O (eds) (2006) Soft Methods for Integrated Uncertainty Modeling. Springer, Berlin

López-Díaz M, Gil MÁ, Grzegorzewski P, Hryniewicz O, Lawry J (eds) (2004) Soft Methodology and Random Information Systems. Springer, Berlin

Taheri SM (2003) Trends in Fuzzy Statistics. Austrian J Stat 32:239–257

Stellar Dynamics, *N*-body Methods for

JUNICHIRO MAKINO
Center for Computational Astrophysics, National Astronomical Observatory of Japan, Tokyo, Japan

Article Outline

Glossary

Binary Two stars which orbit around each other.

Distribution function A density function in the six-dimensional phase space which gives the distribution of stars in a stellar system.

Fokker–Planck equation Partial differential equation for the thermal evolution of the distribution function of a stellar system, expressed in the form of the advection-diffusion equation in the phase space.

Monte-Carlo method The method to numerically follow the thermal evolution of stellar systems using Monte-Carlo integration of Fokker–Planck equation for the distribution function.

Relaxation The process which leads the system to thermal equilibrium. In the case of a stellar system, the main mechanism of relaxation is the gravitational encounter between stars.

Stellar systems A system composed of a large number of stars, i. e., particles which interact through mutual gravity.

Thermal stability Stability of the system against the re-distribution of the thermal energy.

Definition of the Subject

Most of astronomical objects are in the first approximation the collection of point mass particles (stars or planets). We call such systems stellar systems, and stellar dynamics is the theoretical framework for the study of such systems. Since the equation of motion for stellar systems with N stars is analytically solvable only for the case of $N = 2$, numerical integration of the orbit of stars has been very important tool for the study of stellar systems.

Introduction

Stellar Dynamics deals with the evolution of systems which consist of a large number of stars interacting with other stars through mutual gravitational force. Examples of such systems include star clusters, galaxies, clusters of galaxies. Star clusters consist of up to around 10 million stars. They are classified into two categories: open clusters and globular clusters. Open clusters are less massive and younger than globular clusters. Globular clusters in our galaxy are all very old (more than 10 Gyr) and fairly massive (more than 10^5 solar mass). Both types of clusters are compact, with the radius of the order of 10 parsec or less.

Galaxies are larger and more massive than star clusters. The total mass of our Galaxy is around 10^{12} solar mass. Around 80% of the total mass is in dark matter, the existence of which we only indirectly know through the measurement of the motions of visible matter such as gas or stars.

In the first approximation, stellar systems such as star clusters and galaxies are just a collection of a large number of point masses. Therefore, one might expect that a statistical description is applicable. In other words, one might expect that the behavior of the system is described by the statistical mechanics. If that were the case, however, the universe would look very different from its present form. All galaxies would look similar, and rather boring, without any structures such as spirals, bars, shells etc., which make the diversity of galaxies. In this article, we first review how the statistical description can be or cannot be applied to stellar systems, and then present an overview of our current understanding for the evolution of stellar systems. Then we discuss the numerical methods for N-body simulations. For a more complete coverage of the subject see textbooks [5,25].

Statistical Description

Thermal Stability and Gravothermal Catastrophe

There are two basic reasons why stellar systems cannot be described by statistical mechanics. The first one is that the thermal equilibrium state does not exist for stellar systems. An equilibrium state for a system of classical particles is given by the Maxwell-Boltzmann statistics. This means that the local velocity distribution is Maxwellian with a single temperature everywhere in the system. In other words, the velocity distribution function is given by

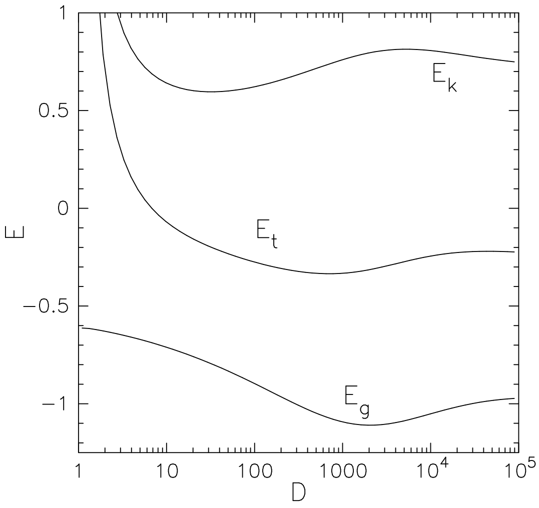
$$f_0(v) = \frac{n_f}{(2\pi\sigma^2)^{3/2}} \exp\left(\frac{-v^2/2}{\sigma^2}\right) , \qquad (1)$$

where n_f is the volume number density of particles and σ is the velocity dispersion.

However, with a stellar system, this Maxwellian distribution cannot be realized, simply because the stars with sufficiently high velocities will escape from the system. The depth of the gravitational potential of a stellar system is finite. Thus, if a star gained kinetic energy higher than that of the depth of the gravitational potential at its position, it would escape from the system.

As a thought experiment, we can think of a stellar system with a spherical adiabatic wall around it. The thermal equilibrium state of the stellar system can be described by that of an ideal gas, since the distribution function is Maxwellian for both cases. Therefore, even though we cannot apply the terms like pressure and temperature to stellar systems in general, we can apply these concepts to stellar systems in the thermal equilibrium, since the equilibrium distribution function is the same as that of an ideal gas with self gravity.

In this case, since the stars do not escape, the equilibrium state does exist. In the limit of very high temperature (large kinetic energy), the distribution of stars within the wall is uniform, with the number density of stars same everywhere. However, as we lower the temperature, the central density becomes higher due to the effect of self gravity. There must be a pressure gradient to support the gravity, and that means a density gradient should exist. However, this increase of density gradient means an increase of the



Stellar Dynamics, *N*-body Methods for, Figure 1
Energies as function of the ratio between the density at the center and that at the place of the wall, *D*. Three curves indicate the kinetic energy (*top*), total energy (*middle*) and gravitational energy (*bottom*), respectively

gravitational force, because more mass is concentrated to the center. Thus, if the increase of the gravitational force is larger than the increase of pressure, than the pressure gradient cannot support the system and there is no equilibrium state.

Figure 1 shows this phenomenon more clearly. Here, we plot the thermal energy (E_k), gravitational energy(E_g) and total energy ($E_t = E_k + E_g$) as a function of the ratio of the density at the center and that of just inside the wall. Here, for simplicity, we use the system of units where the radius of the wall R and mass of the gas M are both unity and the gravitational constant G is also unity. The limit as $D \to 1$ corresponds to infinite temperature. In this limit, the gravitational energy converges to -0.6, and the kinetic energy and total energy diverge. As we increase D, the kinetic energy decreases, gravitational energy increases, and the total energy decreases. However, at $D = 709$, the total energy E_t takes a minimum.

The fact that E_t takes a minimum means that this system has neutral stability against the perturbation which conserves the total energy, since an infinitesimal change of D in either direction does not change the total energy. In other words, for this system we can redistribute the thermal energy in such a way that after that perturbation is applied the system is still isothermal. For $D < 709$, when we change D the total energy also changes. In other words, if we redistribute the energy in any way, as far as we keep the total energy unchanged, the system cannot remain isothermal.

This change of behavior corresponds to a sign of the (second-order) variation of total entropy $\delta^2 S$. For $D < 709$, for any thermal perturbation, $\delta^2 S < 0$, and the second law of thermodynamics guarantees that the isothermal state is stable. However, for $D = 709$, there is one form of perturbation for which $\delta^2 S = 0$. For $D > 709$, there are one or more perturbations for which $\delta^2 S > 0$. In other words, the isothermal state is unstable for $D > 709$.

The eigenfunction which corresponds to $\delta^2 S = 0$ has a single node, which means it either transfers the thermal energy from central area to the outer area, or vise versa. The system shows neutral stability against this heat transfer. Thus, even though the thermal energy is moved, the system remains isothermal.

In the case of $D < 709$, any perturbation in the distribution of the thermal energy results in the heat flux which cancels out the perturbation. In the case of $D = 709$, there is one mode of perturbation for which there will be no heat flux. Finally, in the case of $D > 709$, there is at least one mode of perturbation, for which the resulting heat flows in the direction that enhances perturbation. Thus, if we remove some energy from the outer region and give it to the central region, the central region becomes cooler than the outer region, and therefore heat starts to flow inwards. This inward heat flow results in more cooling.

The reason why the central region cools when it gets more thermal energy is that it expands. As a result of the expansion, the gravitational attraction becomes weaker, and the central region can expand more. Through this additional expansion the central region lowers its temperature. In this case, however, the final state is another isothermal state with the same E_t but a smaller value of D. Since the minimum value of E_t corresponds to $D = 709$, for any other value of D, there is one value of D with same E_t which is thermally stable.

If the initial perturbation removes the thermal energy from the central region and deposits it to the outer region, the central region becomes hotter, and the heat starts to flow outwards. In this case, there is no stable final state, and the central region continues to contract and becomes hotter indefinitely. Exactly what happens depends on the efficiency of the heat transport.

Roughly speaking, if the thermal timescale is shorter in the center than in the outer region, as the central region contracts, the timescale becomes even shorter. In this case, the central density can reach infinity in a finite time. Of course, in this case, the mass of the region with infinite density should become zero. On the other hand, if the timescale is longer in the center, the timescale becomes even longer as the contraction proceeds, and the overall system contracts slowly. In the case of an ideal gas, the timescale is determined by the mechanism of the heat transfer.

In the case of stellar systems, heat transfer proceeds through close encounters between two stars. Since the rate of close encounters is directly proportional to the number density of stars, the thermal relaxation timescale is shorter for the central region, and becomes shorter as the central density goes up. This phenomenon is called "gravothermal catastrophe".

This gravothermal catastrophe occurs in real stellar systems without the adiabatic wall, since real stellar systems always have the temperature decreasing outward, which naturally drives the outward heat flux. Of course, the distribution function is not Maxwellian and one cannot strictly define the temperature. Even so, comparisons between numerical simulations with different approaches have shown that this picture gives a very good description.

Hachisu et al. [9] performed a simulation of the evolution of the self-gravitating gas system with several different forms of heat conductivity. They found that if the heat conductivity is such that the central thermal timescale is shorter than that of the outside region, the evolution of the system would become self-similar. In other words, the central region continues to shrink, leaving a power-law halo behind it. The mass of the core decreases in time and the central density reaches infinity within a finite time.

In the following, we give a simple explanation of this power-law behavior, following the description by Lynden-Bell and Eggleton [14]. Formally, a self-similar solution for a physical quantity y as a function of radius r and time t is expressed as

$$y(r,t) = y_0(t)y_*[r/r_0(t)] . \tag{2}$$

We can set $y_*(0) = 1$ without loss of generality. We can assume that in the limit of $r \to \infty$ there is no evolution, since the thermal timescale is longer at the outskirts. We can further assume that functions r_0 and y_0 are powers of the time t, since otherwise the self-similar solution cannot be constructed. Thus, if we express

$$r_0 = (t_0 - t)^\beta , \tag{3}$$

and

$$y_0 = (t_0 - t)^\gamma , \tag{4}$$

we have

$$y_0 = r_0^{\gamma/\beta} . \tag{5}$$

The ratio between the gravitational binding energy of the core and the thermal energy of the core should be constant. Therefore, we have

$$\sigma^2 \propto \frac{GM_c}{r_c} \sim \rho_0 r_0^2 . \tag{6}$$

If we express ρ_0 as

$$\rho_0 = r_0^\alpha , \tag{7}$$

we have

$$r_0 = (t_0 - t)^{2/(6+\alpha)} . \tag{8}$$

Lynden-Bell and Eggleton [14] numerically obtained the self-similar solution for the gaseous model with heat conductivity which they believed would mimic the radial energy transfer in an N-body system, and found that the self-similar solution has the characteristic power law index of

$$\rho = r^{-2.21} . \tag{9}$$

Henon [11] demonstrated that the N-body system would exhibit core collapse using Monte-Carlo calculation with 1000 shells. At that time, N-body simulation did show some collapse-like behavior, but it was difficult to see whether the collapse is really self-similar or not because of the limitation in the number of particles. The most beautiful demonstration of the self-similar nature of the collapse is by Cohn [6], who is the first to use the direct integration of the Fokker–Planck equation in the study of the thermal evolution of the globular clusters. He found the power index to be −2.23, which is strikingly close to the value obtained by the gas model calculation of Lynden-Bell and Eggleton [14].

In the study of the gravothermal catastrophe and self-similar solution, N-body simulation did not play a major role. This is partly because more approximate methods, such as the gaseous models and Fokker–Planck calculations, gave reasonable results, and partly because computer power available was quite limited in the 1980s.

Binary Formation and Gravothermal Oscillation

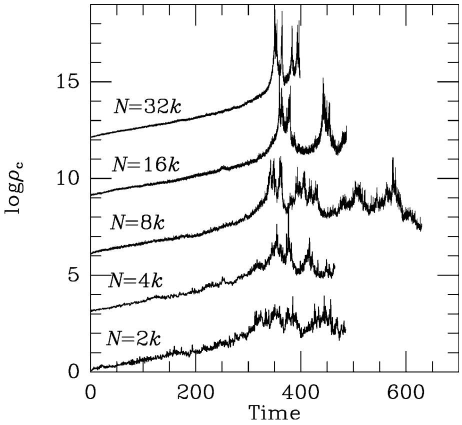
In the continuous limit, the central density diverges in a finite time, and at that time the core mass becomes zero. Since real stellar systems consist of stars with finite mass and finite size, the central density cannot reach infinity. One possibility is that stars start to physically collide, but there is another possibility.

When the central density becomes very high, the cross-section of the three-body close encounter, which causes the formation of a bound binary, becomes non-negligible.

The binding energy of the binary is transferred to the kinetic energy of the third particle and the center-of-mass motion of the binary. This is essentially the same as the nuclear fusion reaction. The difference is that in this case the interaction which makes the two stars bound is the same gravitational force which makes the entire system bound, and that means there is no ground state for a binary star. Thus, in theory, the binding energy of a binary star can become arbitrarily large. In practice, in real star clusters, there is a practical limit for the binding energy of a binary star. When a binary star interacts with another star, its center-of-mass motion acquires some energy, which is typically a fraction of its binding energy. When the kinetic energy of the center-of-mass motion becomes large enough, the binary would be ejected out the cluster. In fact, this ejection of the binary itself and the other star is the main channel which heats the central region of a star cluster.

This energy production by binaries halts the collapse. The collapse of the core is driven by the outward heat flux. If a sufficient amount of energy is generated in the core, the collapse is halted. In the case of a star, gravitational contraction of the star is halted by the energy production from the nuclear fusion reaction, and it becomes a main-sequence star.

In the case of a star cluster, it is possible to construct a steady-state solution quite similar to the main-sequence stage of a star [8,10]. However, it turned out that this steady-state solution is again thermally unstable [26]. This instability drives what is now called "gravothermal oscillation". This oscillation was first found in a gas sphere model, but then confirmed with Fokker–Planck (FP) calculations [7], and later with direct *N*-body simulations [16].



Stellar Dynamics, *N*-body Methods for, Figure 2
The time variation of the central density of simulated star clusters. (Reproduced from [16])

Figure 2 shows the time variation of the central density for simulations with 2–32k particles. The time is scaled so that the initial thermal relaxation time is the same for all runs. The curves are shifted vertically.

From Fig. 2 we can see that the oscillation is not strictly periodic. One reason is simply that the binary formation and its interaction with other stars is a stochastic process. The other reason is that, even in the continuous limit with a smooth and deterministic heating term, this oscillation is chaotic. Note that here what exhibits the chaotic behavior is the central density of the whole *N*-body system, which is a macroscopic variable of a system with large degrees of freedom.

Numerical Methods

In this section, we discuss the numerical methods used to obtain the solution of gravitational *N*-body problem. The problem itself is rather simple. We integrate the equation of motion for stars, using some numerical methods for the initial-value problem of ordinary differential equations. The calculation cost of the mutual gravitation interaction is $O(N^2)$ and dominate the calculation cost.

In practice, however, what is done is far more complex for a variety of reasons. Here, let us overview the approaches used.

First of all, for some problems, it is not necessary to calculate the interaction of all $N^2/2$ pairs at each timestep. There are several algorithms which reduce the calculation cost from $O(N^2)$ to $O(N \log N)$ or even $O(N)$. These algorithms include particle-mesh schemes, Barnes-Hut tree algorithm, and the Fast Multipole Method.

Secondly, it is not practical to apply the same timestep to all stars. While two stars undergo close encounters, their distance can become arbitrary small, depending on the impact parameter and the relative velocity. They have to be integrated with the adaptive timestep which can resolve their relative orbit. Moreover, as we discussed in Sect. "Statistical Description", many stellar systems tend to develop high-density cores, in which the stars orbit on a timescale much shorter than that of average stars. Also, in many stellar systems, interactions between binary stars (two stars orbiting around each other) and other stars play an important role, and the evolution of the binary orbit due to the interaction with other stars must be accurately integrated. The orbital timescale of binaries can become as small as a fraction of a second, while stars in globular clusters or galaxies have the orbital timescale much longer than one million years. Thus, it is clearly necessary to apply some algorithm with which we can change the timestep of individual stars individually.

Finally, an important characteristic of a stellar system is, from both the theoretical and practical point of views, that it is a Hamiltonian system with various symmetry and conservation laws. Numerical schemes which make use of these characteristics play important roles.

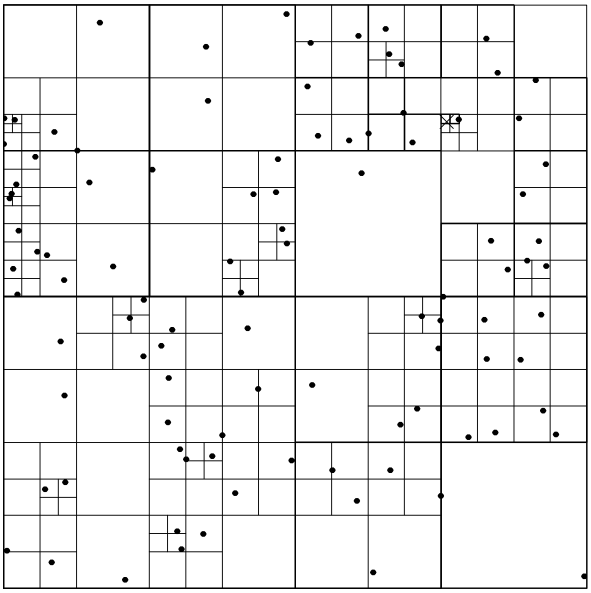
In the following, we briefly overview these schemes.

Fast Interaction Calculations

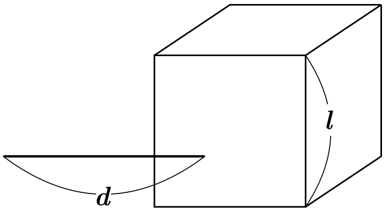
The basic idea of the tree code [4] is to replace the force from a group of distant particles with the force from their center of mass or by a multipole expansion. To ensure accuracy, we make groups for distant particles large and groups for nearby particles small.

We use a tree structure to construct the appropriate grouping for each particle. Before calculating the forces on particles, we first organize particles into a tree structure. Barnes and Hut [4] used an oct-tree based on the recursive subdivision of a cube into eight subcubes. We stop the recursive subdivision if the cube has only one particle (or no particles). See [15] for details concerning an efficient tree construction algorithm. Figure 3 shows the Barnes–Hut tree in two dimensional space.

After the tree is constructed, for each node of the tree, which corresponds to a cube of a certain size, we calculate the coefficient of the multipole expansion of the gravitational force exerted by particles in that cube. The fast algorithm was described in [12].



Stellar Dynamics, *N*-body Methods for, Figure 3
Barnes–Hut tree in two dimensions



Stellar Dynamics, *N*-body Methods for, Figure 4
Opening criterion for tree traversal

The force calculation is expressed as a recursive procedure. To calculate the force on a particle, we start from the root node, which corresponds to the total system. We calculate the distance between the node and the particle (d) and compare it with the size of the node (l). If they satisfy the convergence criterion

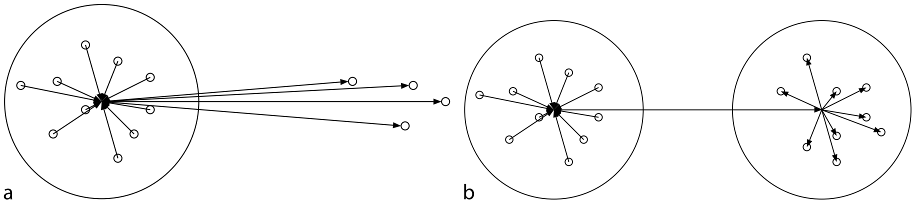
$$\frac{l}{d} < \theta \,, \tag{10}$$

where θ is the accuracy parameter, we calculate the force from that node to the particle using the coefficients of the multipole expansion. If criterion (10) is not satisfied, the force is calculated as a summation of the forces from eight sub-nodes.

Usually, we use the distance between the particle and the center of mass of the node to determine whether the force is accurate enough. When θ is very large, this criterion can cause unacceptably large error [21]. For most calculations, however, such a pathological situation is not realized.

The fast multipole algorithm (FMM) is in some sense a natural extension of the tree algorithm. In the tree algorithm, we use the multipole expansion to express the gravitational force for a group of stars with other stars sufficiently far away (Fig. 5a). In FMM, the forces on another group of stars are expressed in terms of one spherical harmonics expansion, and the forces on these stars are obtained by evaluating the expansion at the locations of stars. The tree-based recursive approach is now applied to both the path to calculate multipole expansions and the path to calculate the spherical harmonics expansions (usually called local expansions). This treatment removes the log N term in the calculation cost of the tree algorithm, since the cells of a given size do not directly interact with stars. They interact only with the cells of the same or similar sizes.

After the invention of the tree algorithm and FMM in the late 1980s, a large number of works followed on various topics such as the theoretical analysis of the algo-



Stellar Dynamics, *N*-body Methods for, Figure 5
Tree algorithm (a) and FMM (b)

rithms, efficient implementation, a fast method to translate multipole and local expansions, parallel implementation, etc. Some of these are covered in [19].

In many other fields of computational science, the Poisson equation associated with the distribution of particles is numerically solved using FFT. If the number of grid points one needs to resolve the potential field is of the same order or less than the number of particles, it is highly advantageous to use FFT, since the calculation cost of FFT is $O(M \log M)$, where M is the number of grid points, with relatively small coefficient. However, FFT requires that the grid is regular with equal spacing. Thus, FFT alone is not very suitable for astronomical simulations, where small-scale structures develop through gravitational instability. Recently, the combination of the tree algorithm and FFT potential solver has become the standard scheme for simulations which requires periodic boundary conditions (see, e. g., Yoshikawa and Fukushige [27]).

Individual Timestep Algorithms

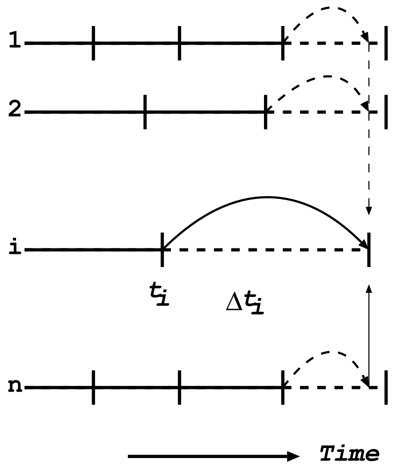
The individual timestep scheme [1,2] has been the only algorithm that can be used for simulations of gravitational many-body systems, such as open clusters, globular clusters and a system of planetesimals. In a simulation of these systems, we are interested in the change of orbit of each particle due to gravitational encounters with other particles, and the evolution of the total system driven by such changes. In that sense, simulation of these systems is similar to the molecular dynamics simulation, in which we are interested in the thermodynamical process.

In simulations of these collisional systems, we need to follow the changes of the orbit of a star due to individual encounters with a reasonable accuracy, since the encounters drive the evolution of the system. Therefore, we cannot use the softening parameter to simulate the evolution of these systems. In the study of the galaxies, for example, we can modify the $1/r$ potential to $1/\sqrt{r^2 + \epsilon^2}$, with small constant ϵ. This softened potential makes it possible to use a constant and global timestep. When we use a strict $1/r$ potential, the force between two particles changes very rapidly when two particles undergo a close encounter. Therefore, during a close encounter, the timestep should be sufficiently small to resolve this rapid change. Roughly speaking, the timestep is determined by the distance to the nearest neighbor. Therefore, if we integrate the system in lockstep, the timestep is determined by the pair of particles with minimum separation. Even in a nearly homogeneous system, the minimum separation is proportional to $N^{-2/3}$, and the dependence on N is stronger when the system is highly inhomogeneous [17].

To reduce the total calculation cost, Aarseth [1] developed an algorithm which assigns each particle its own timestep. In this individual timestep algorithm, each particle adjusts its timestep so that it satisfies the required accuracy. Thus, when two particles undergo a close encounter, although the timesteps of these particles shrink as required, the timesteps of other particles remain long. Makino and Hut [17] showed that the individual timestep scheme is faster than a shared timestep scheme by a factor in the range of $O(N^{1/3})$ and $O(N)$, depending on the distribution of particles.

In the individual timestep algorithm, each particle has its own timestep, and therefore its own time. To calculate the force on a particle due to other particles, we must know their positions at the time of the particle for which we calculate the force. To calculate the position of a particle at a time different from the time of the particle, Aarseth used a third order polynomial extrapolation. This polynomial is evaluated each time a pairwise force is calculated, therefore the calculation cost of the polynomial evaluation is of the same order as that of the force calculation itself. The basic algorithm looks like the following in the case of Aarseth's program, which uses a predictor-corrector scheme:

(a) Select particle i with a minimum $t_i + \Delta t_i$. Set the global time (t) to be this minimum, $t_i + \Delta t_i$.
(b) Predict the positions of all the particles at time t using the extrapolation polynomial.
(c) Calculate the acceleration (a_i) for particle i at time t, using the predicted positions.
(d) Apply the corrector for the position and velocity of particle i.
(e) Go back to step (a).



Stellar Dynamics, *N*-body Methods for, Figure 6
Schematic description of the individual timestep algorithm

Figure 6 illustrates how the individual timestep scheme works. When particle i is integrated from t_i to $t_i + \Delta t_i$, the positions of all other particles are predicted at that time, and forces from those particles are calculated using these predicted positions.

Not all time integration schemes can be used with this individual timestep, since the predicted values of the positions of all other particles need to be calculated at the time of the particle to be integrated. Most integration schemes do not provide the solutions at arbitrary points in time. For example, there is no simple way to obtain the approximate solution at the intermediate time with the usual Runge–Kutta schemes or extrapolation schemes.

For the treatment of binaries and small-N systems which can formed in larger stellar systems, see Aarseth [3].

Special Numerical Methods

A stellar system is a Hamiltonian system, and therefore the symplectic algorithms [22] can in principle be applied. These schemes show behavior much better than that of traditional schemes, and can dramatically reduce the integration error for long calculations. However, in practice there are a number of difficulties. One is that with symplectic schemes we cannot change timesteps [24]. There has been a number of works to make the symplectic scheme effectively work as variable-timestep scheme. Most of them rely on the partition of the potential energy term of the Hamiltonian into multiple parts with different timescales, and apply different timesteps to different parts of the potential term [20,23]. A very different approach is to use time-symmetric scheme instead of symplectic schemes. Symmetric schemes offer most of the practical advantages of symplectic schemes, and remain symmetric if we change the timestep. Here, the timestep must be calculated in such a way that does not destroy the time symmetry [13]. Recently, a way to combine this symmetric algorithm and the individual timestep algorithm was proposed [18].

Future Directions

In Sect. "Statistical Description" we over viewed the evolution of a system of point-mass particles. When we try to understand the evolution of real stellar systems such as globular clusters, open clusters and galactic nuclei, we need to take into account various effects which were neglected in Sect. "Statistical Description". For example, real stars are not point masses but have finite radii, and can physically collide with each other. In the case of normal stars like our sun in a galaxy, the chance that it will collide with its neighbors before the end of its lifetime is negligible. However, the number density of stars in the central regions of globular clusters or other massive star clusters can reach more than $10^6/\mathrm{pc}^3$, or about a million times that of the solar neighborhood. In such circumstances the collision is not rare. Also, individual stars have different masses. Even in the case of thermal equilibrium, stars with different masses have different spacial distribution, since the kinetic energy per unit mass is different. Thus, massive stars tend to lose their kinetic energy and segregate to the central region of the cluster, resulting in strong enhancement of collision rates between most massive stars in star clusters. To make things even more complex, each star evolves from the Main Sequence to giant, and depending on their masses, goes through Type II supernova and finally becomes a white dwarf, a neutron star, or a black hole. These stellar evolutions and stellar collisions, coupled with the stellar dynamics, makes stellar systems very interesting and an important research subject.

The ultimate research direction is to model all of these processes, stellar evolution, stellar collisions, stellar dynamics, and at some point the initial gas dynamics of star cluster formation, in one single unified simulation. From the viewpoint of the computational cost, such a simulation is not impossible, since the calculation cost of stellar dynamics is high and others are relatively minor. However, from the viewpoint of software engineering and physical modeling, we do not really know how we can actually develop and maintain a simulation program which can handle gravitational interaction between stars, evolution of individual stars, physical collisions and tidal interaction between stars, and other more complex things like the evolution of binary stars. MODEST (Modeling Dense Stellar systems)[1] is one of such efforts to develop a simulation code, or a loosely coupled collection of codes, to handle complex systems.

Bibliography

1. Aarseth SJ (1963) MN 126:223
2. Aarseth SJ (1985) In: Blackbill JU, Cohen BI (eds) Multiple Time Scales. Academic Press, New York, pp 377–418
3. Aarseth SJ (2003) Gravitational N-Body Simulations. Cambridge University Press, Cambridge, pp 430
4. Barnes J, Hut P (1986) Nature 324:446
5. Binney J, Tremaine S (2008) Galactic Dynamics, 2nd edn. Princeton University Press, Princeton
6. Cohn H (1980) ApJ 242:765
7. Cohn H, Hut P, Wise M (1989) ApJ 342:814
8. Goodman J (1984) ApJ 280:298
9. Hachisu I, Nakada Y, Nomoto K, Sugimoto D (1978) Prog Theor Phys 60:393
10. Heggie DC (1984) MN 206:179
11. Henon M (1971) Astrophys Space Sci 13:284
12. Hernquist L (1990) J Comp Phys 87:137
13. Hut P, Makino J, McMillan S (1995) ApJL 443:L93
14. Lynden-Bell D, Eggleton PP (1980) MN 191:483
15. Makino J (1990) J Comput Phys 87:148
16. Makino J (1996) Dynamical Evolution of Star Clusters. In: Hut P, Makino J (eds) Proc IAU Symp 174. Kluwer, Amsterdam, pp 151–160
17. Makino J, Hut P (1988) ApJS 68:833
18. Makino J, Hut P, Kaplan M, Saygın H (2006) New Astron 12:124
19. Pfalzner S, Gibbon P (1996) Many-Body Tree Methods in Physics. Cambridge University Press, Cambridge
20. Preto M, Tremaine S (1999) AJ7 118:2532
21. Salmon JK, Warren MS (1994) J Comp Phys 111:136
22. Sanz-Serna JM, Calvo MP (1994) Numerical Hamiltonian Problems. Chapman and Hall, London
23. Skeel RD, Biesiadecki JJ (1994) Ann Numer Math 1:191
24. Skeel RD, Gear CW (1992) Physica D 60:311
25. Spitzer Jr L (1987) Dynamical Evolution of Globular Clusters. Princeton University Press, Princeton
26. Sugimoto D, Bettwieser E (1983) MN 204:19P
27. Yoshikawa K, Fukushige T (2005) PASJ 57:849

Stochastic Games

Eilon Solan
The School of Mathematical Sciences,
Tel Aviv University, Tel Aviv, Israel

Article Outline

Glossary

A stochastic game A repeated interaction between several participants in which the underlying state of the environment changes stochastically, and it depends on the decisions of the participants.

A strategy A rule that dictates how a participant in an interaction makes his decisions as a function of the observed behavior of the other participants and of the evolution of the environment.

Evaluation of stage payoffs The way that a participant in a repeated interaction evaluates the stream of stage payoffs that he receives (or stage costs that he pays) along the interaction.

An equilibrium A collection of strategies, one for each player, such that each player maximizes (or minimizes, in case of stage costs) his evaluation of stage payoffs given the strategies of the other players.

A correlated equilibrium An equilibrium in an extended game in which at the outset of the game each player receives a private signal, and the vector of private signals is chosen according to a known joint probability distribution. In the extended game, a strategy of a player depends, in addition to past play, on the signal he received.

Definition of the Subject

Stochastic games, first introduced by Shapley [60], model dynamic interactions in which the environment changes in response to the behavior of the players. Formally, a stochastic game is a tuple $G = \langle N, S, (\mathcal{A}_i, A_i, u_i)_{i \in N}, q \rangle$ where

- N is a set of players.
- S is a state space. If S is uncountable, it is supplemented with a σ-algebra of measurable sets.
- For every player $i \in N$, $\mathcal{A}_i$ is a set of actions for that player, and $A_i \colon S \to \mathcal{A}_i$ is a set-valued (measurable) function that assigns to each state $s \in S$ the set of actions $A_i(s)$ that are available to player i in state s. If $\mathcal{A}_i$ is uncountable, it is supplemented with a σ-algebra of measurable sets. Denote $SA = \{(s, a) \colon s \in S, a =$

[1] http://www.manybody.org/modest/

$(a_i)_{i \in N}, a_i \in A_i(s) \;\; \forall i \in N\}$. This is the set of all possible *action profiles*.

- For every player $i \in N$, $u_i : SA \to \mathbf{R}$ is a (measurable) *stage payoff function* for player i.
- $q : SA \to \Delta(S)$ is a (measurable) *transition function*, where $\Delta(S)$ is the space of probability distributions over S.

The game starts at an initial state s^1, and is played as follows. At each stage $t \in \mathbf{N}$, each player $i \in N$ chooses an action $a_i^t \in A_i(s^t)$, receives the stage payoff $u_i(s^t, a^t)$, where $a^t = (a_i^t)_{i \in N}$, and the game moves to a new state s^{t+1} that is chosen according to the probability distribution $q(\cdot \mid s^t, a^t)$.

A few comments are in order.

1. A stochastic game lasts infinitely many stages. However, the model also captures finite interactions (of length t), by assuming the play moves, at stage t, to an absorbing state with payoff 0 to all players.
2. In particular, by setting $t = 1$, we see that stochastic games are a generalization of matrix games (games in normal formgames in normal form), which are played only once.
3. Stochastic games are also a generalization of repeated games, in which the players play the same matrix game over and over again. Indeed, a repeated game is equivalent to a stochastic game with a single state.
4. Stopping games are also a special case of stochastic games. In these games every player has two actions in all states, *continue* and *quit*. as long as all players choose *continue* the stage payoff is 0; once at least one player chooses *quit* the game moves to an absorbing state.
5. Markov decision problems (see, e. g., [49]) are stochastic games with a single player.
6. The transition function q governs the evolution of the game. It depends on the actions of all players and on the current state, so that all the players influence the evolution of the game.
7. The payoff function u_i of player i depends on the current state as well as on the actions chosen by all players. Thus, a player's payoff depends not only on that player's choice, but also on the behavior of the other players.
8. Though we refer to the functions $(u_i)_{i \in N}$ as "stage payoffs", with the implicit assumption that each player tries to maximize his payoff, in some applications these functions describe a stage *cost*, and then the implicit assumption is that each player tries to minimize his cost.
9. The action of a player at a given stage affects both his stage payoff and the evolution of the state variable, thereby affecting his future payoffs. These two, sometimes contradicting effects make the optimization problem of the players quite intricate, and the analysis of the game challenging.
10. The players receive a stage payoff at each stage. So far we did not mention how the players evaluate the infinite stream of stage payoffs that they receive, nor did we say what is their information at each stage: Do they observe the current state? Do they observe the actions of the other players? These issues will be discussed later.
11. The actions that are available to the players at each stage, the payoff functions, and the transition function, all depend on the current state, and not on past play (that is, past states that the game visited, and past actions that the players chose). This assumption is without loss of generality. Indeed, suppose that the actions available to the players at each stage, the payoff functions, and the transition function, all depend on past play, as well as on the current state. For every $t \in \mathbf{N}$ let H_t be the set of all possible *histories* of length t, that is, all sequences of the form $(s^1, a^1, s^2, a^2, \dots, s^t)$, where $s^k \in S$ for every $k = 2, 3, \dots, t$, $a^k = (a_i^k)_{i \in N}$ and a_i^k is an available action to player i at stage k, for every $k = 1, 2, \dots, t-1$. Then the game is equivalent to a game with state space $H := \bigcup_{t \in \mathbf{N}} H_t$, in which the state variable captures past play, and the state at stage t lies in H_t. In the new game, the sets of available actions, the payoff function, and the transition function, depend on the current state rather than on all past play.

The interested reader is referred to [20,42,72] for further reading on stochastic games. We now provide a few applications.

Example 1 (Capital Accumulation ([7,18,19,34,45])) Two (or more) agents jointly own a natural resource or a productive asset; at every period they have to decide the amount of the resource to consume. The amount that is not consumed grows by a known (or an unknown) fraction. Such a situation occurs, e. g., in fishery: Fishermen from various countries fish in the same area, and each country sets a quota for its fishermen. Here the state variable is the current amount of resource, the action set is the amount of resource to be exploited in the current period, and the transition is influenced by the decisions of all the players, as well as possibly by the random growth of the resource.

Example 2 (Taxation ([14,48])) A government sets a tax rate at every period. Each citizen decides at every period how much to work, and, from the total amount of money he or she has, how much to consume; the rest is saved for the next period, and grows by a known interest rate. Here the state is the amount of savings each citizen has, the stage payoff of a citizen depends on the amount of money that he consumed, on the amount of free time he has, and on the total amount of tax that the government collected. The stage payoff of the government may be the average stage payoff of the citizens, the amount of tax collected, or a mixture of the two.

Example 3 (Communication Network [58]) A single-cell system with one receiver and multiple uplink transmitters share a single, slotted, synchronous classical collision channel. Assume that all transmitted packets have the same length, and require one time unit, which is equal to one time slot, for transmission. Whenever a collision occurs, the users attempt to retransmit their packets in subsequent slots to resolve collision for reliable communication.

Here a state lists all relevant data for a given stage: e. g., the number of packets waiting at each transmitter, or the length of time each has been waiting to be transmitted. The players are the transmitters, and the action of each transmitter is which packet to transmit, if any. The stage cost may depend on the number of time slots that the transmitted packet waited, on the number of packets that have not been transmitted at that period, and possibly on additional variables. The transition depends on the actions chosen by the players, but it has a stochastic component, which captures the number of new packets that arrive at the various transmitters during every time slot.

Example 4 (Queues [1]) Individuals that require service have to choose whether to be served by a private slow service provider, or by a powerful public service provider. This situation arises, e. g., when jobs can be executed on either a slow personal computer or a fast mainframe. Here a state lists the current load of the public and private service providers, and the cost is the time to be served.

The importance of stochastic games stems from the wide range of applications they encompass. Many repeated interactions can be recast as stochastic games; the wide range of theoretical results that have been obtained provide insights that can help in analyzing specific situations and suggesting proper behavior to the participants. In certain classes of games algorithms that have been developed may be used to calculate such behavior.

Strategies, Evaluations and Equilibria

So far we have not described the information that the players have at each stage. In most of the chapter we assume that the players have complete information of past play; that is, at each stage t, they know the sequence $s^1, a^1, s^2, a^2, \dots, s^t$ of states that were visited in the past (including the current state) and the actions that were chosen by all players. This assumption is too strong for most applications, and in the sequel we will mention the consequences of its relaxation.

Since the players observe past play, a *pure strategy* for player i is a (measurable) function σ_i that assigns to every finite history $(s^1, a^1, s^2, a^2, \dots, s^t)$ an action $\sigma_i(s^1, a^1, s^2, a^2, \dots, s^t) \in A_i(s^t)$, with the interpretation that, at stage t, if the finite history $(s^1, a^1, s^2, a^2, \dots, s^t)$ occurred, player i plays the action $\sigma_i(s^1, a^1, s^2, a^2, \dots, s^t)$. If the player does not know the complete history, then a strategy for player i is a function that assigns to every possible information set, an action that is available to the player when the player has this information. A *mixed strategy* for player i is a probability distribution over the set of his pure strategies. The space of mixed strategies of player i is denoted by σ_i.

A simple class of strategies is the class of *stationary* strategies; a strategy σ_i for player i is *stationary* if $\sigma_i(s^1, a^1, s^2, a^2, \dots, s^t)$ depends only on the current state s^t, and not on past play $s^1, a^1, s^2, a^2, \dots, a^{t-1}$. A stationary strategy of player i can be identified with an element $x = (x_s)_{s\in S} \in \times_{s\in S} \Delta(A_i(s))$, with the interpretation that player i plays the mixed action x_s whenever the current state is s. Denote by $X_i = \times_{s\in S} \Delta(A_i(s))$ the space of stationary strategies of player i.

There are three common ways to evaluate the infinite stream of payoffs that the players receive in a stochastic game: The *finite-horizon evaluation*, in which a player considers the average payoff during the first T stages, the *discounted evaluation*, in which a player considers the discounted sum of his stage payoffs, and the *limsup evaluation*, in which a player considers the limsup of his long-run average payoffs. We now formally define these evaluations.

Every profile $\sigma = (\sigma_i)_{i\in N}$ of mixed strategies, together with the initial state, induces a probability distribution $\mathbf{P}_{s_1,\sigma}$ over the space of infinite plays $H_\infty := SA^{\mathbf{N}}$. We denote the corresponding expectation operator by $\mathbf{E}_{s_1,\sigma}$.

Definition 5 Let σ be a profile of mixed strategies. For every finite horizon $T \in \mathbf{N}$, the *T-stage payoff* under σ for player i is

$$\gamma_i^T(s_1, \sigma) := \mathbf{E}_{s_1,\sigma}\left[\frac{1}{T}\sum_{t=1}^{T} u_i(s^t, a^t)\right] .$$

For every discount factor $\lambda \in (0, 1]$, the λ-*discounted payoff* under σ for player i is

$$\gamma_i^\lambda(s_1, \sigma) := \mathbf{E}_{s_1,\sigma}\left[\lambda \sum_{t=1}^{\infty}(1-\lambda)^{t-1} u_i(s^t, a^t)\right] .$$

The *limsup payoff* under σ for player i is

$$\gamma_i^\infty(s_1, \sigma) := \mathbf{E}_{s_1,\sigma}\left[\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} u_i(s^t, a^t)\right] .$$

The T-stage payoff captures the situation in which the interaction lasts exactly T stages. The λ-discounted evaluation captures the situation in which the game lasts "many" stages, and the player discounts stage payoffs – it is better to receive \$1 today than tomorrow. The limsup payoff also captures the situation in which the game lasts "many" stages, but here the player does not discount his payoffs, and the payoff at each given stage is insignificant as compared to the payoff in all other stages. Equivalently, one could consider the liminf payoff in which the player considers the liminf of the long-run average payoffs.

As usual, an equilibrium is a vector of strategies such that no player can profit by a unilateral deviation. For every player i and every strategy profile $\sigma = (\sigma_i)_{i\in N}$ we denote the strategy profile of all other players, except player i, by $\sigma_{-i} = (\sigma_j)_{j\neq i}$.

Definition 6 Let $\varepsilon \geq 0$. A profile of strategies σ is a T-*stage* ε-*equilibrium* if

$$\gamma_i^T(s_1, \sigma) \geq \gamma_i^T(s_1, \sigma_i', \sigma_{-i}) - \varepsilon , \quad \forall s_1 \in S, \forall i \in N, \forall \sigma_i' \in \Sigma_i .$$

It is a λ-*discounted* ε-*equilibrium* if

$$\gamma_i^\lambda(s_1, \sigma) \geq \gamma_i^\lambda(s_1, \sigma_i', \sigma_{-i}) - \varepsilon , \quad \forall s_1 \in S, \forall i \in N, \forall \sigma_i' \in \Sigma_i .$$

It is a *limsup* ε-*equilibrium* if

$$\gamma_i^\infty(s_1, \sigma) \geq \gamma_i^\infty(s_1, \sigma_i', \sigma_{-i}) - \varepsilon , \quad \forall s_1 \in S, \forall i \in N, \forall \sigma_i' \in \Sigma_i .$$

The payoff that corresponds to an ε-equilibrium, that is, either one of the quantities $\gamma^T(s_1, \sigma)$, $\gamma^\lambda(s_1, \sigma)$ and $\gamma^\infty(s_1, \sigma)$, is called an ε-*equilibrium payoff* at the initial state s_1.

As we will see below, when both state and action spaces are finite, a T-stage and a λ-discounted 0-equilibrium exist. However, when the state or action spaces are infinite such a 0-equilibrium may fail to exist, yet ε-equilibria may exist for every $\varepsilon > 0$.

As the length of the game T varies, or as the discount factor λ varies, the equilibrium strategy profile varies as well. A strategy profile that is an ε-equilibrium for every T sufficiently large and for every λ sufficiently small is called a *uniform* ε-*equilibrium*.

Definition 7 Let $\varepsilon > 0$. A strategy profile σ is a *uniform* ε-*equilibrium* if there are $T_0 \in \mathbf{N}$ and $\lambda_0 \in (0, 1)$ such that for every $T \geq T_0$ the strategy profile σ is a T-stage ε-equilibrium, and for every $\lambda \in (0, \lambda_0)$ it is a λ-discounted ε-equilibrium.

If for every $\varepsilon > 0$ the game has a (T-stage, λ-discounted, limsup or uniform) ε-equilibrium with corresponding payoff g_ε, then any accumulation point of $(g_\varepsilon)_{\varepsilon>0}$ as ε goes to 0 is a (T-stage, λ-discounted, limsup or uniform) equilibrium payoff.

Zero-Sum Games

A two-player stochastic game is *zero-sum* if $u_1(s, a) + u_2(s, a) = 0$ for every $(s, a) \in SA$. As in matrix games, every two-player zero-sum stochastic game admits at most one equilibrium payoff at every initial state s_1, which is termed the *value* of the game at s_1. Each player's strategy which is part of an ε-equilibrium is termed ε-*optimal*. The definition of ε-equilibrium implies that an ε-optimal strategy guarantees the value up to ε; for example, in the T-stage evaluation, if σ_1 is an ε-optimal strategy of player 1, then for every strategy of player 2 we have

$$\gamma_1^T(s_1, \sigma_1, \sigma_2) \geq v^T(s_1) - \varepsilon ,$$

where $v^T(s_1)$ is the T-stage value at s_1.

In his seminal work, Shapley [60] presented the model of two-player zero-sum stochastic games with finite state and actions spaces, and proved the following.

Theorem 8 [60] *For every two-player zero-sum stochastic game, the λ-discounted value at every initial state exists. Moreover, both players have λ-discounted 0-optimal stationary strategies.*

Proof Let $\mathcal{V}$ be the space of all functions $v\colon S \to \mathbf{R}$. For every $v \in \mathcal{V}$ define a zero-sum matrix game $G_s^\lambda(v)$ as follows:

- The action spaces of the two players are $A_1(s)$ and $A_2(s)$ respectively.
- The payoff function (that player 2 pays player 1) is

$$\lambda u_1(s, a) + (1-\lambda)\sum_{s'\in S} q(s' \mid s, a) v(s') .$$

The game $G_s^\lambda(v)$ captures the situation in which, after the first stage, the game terminates with a terminal payoff $v(s')$, where s' is the state reached after stage 1. Define an operator $\varphi\colon \mathcal{V} \to \mathcal{V}$ as follows:

$$\varphi_s(v) = \mathrm{val}(G_s^\lambda(v)) \, ,$$

where $\mathrm{val}(G_s^\lambda(v))$ is the value of the matrix game $G_s^\lambda(v)$. Since the value operator is non-expansive, it follows that the operator φ is contracting: $\|\varphi(v) - \varphi(w)\|_\infty \leq (1 - \lambda)\|v - w\|_\infty$, so that this operator has a unique fixed point $\widehat{v}^\lambda$. One can show that the fixed point is the value of the stochastic game, and every strategy σ_i of player i in which he plays, after each finite history $(s^1, a^1, s^2, a^2, \ldots, s^t)$, an optimal mixed action in the matrix game $G_{s^t}^\lambda(\widehat{v}^\lambda)$, is a λ-discounted 0-optimal strategy in the stochastic game. □

Example 9 Consider the following two-player zero-sum game with three states, s_0, s_1 and s_2; each entry of the matrix indicates the payoff that player 2 (the column player) pays player 1 (the row player, the payoff is in the middle), and the transitions (which are deterministic, and are denoted at the top-right corner).

$$\begin{array}{c|c|c|} & L & R \\ \hline T & 0 \;\; s^2 & 1 \;\; s^1 \\ \hline B & 1 \;\; s^1 & 0 \;\; s^0 \\ \hline \end{array} \qquad \begin{array}{c|c|} & L \\ \hline T & 1 \;\; s^1 \\ \hline \end{array} \qquad \begin{array}{c|c|} & L \\ \hline T & 0 \;\; s^0 \\ \hline \end{array} \; .$$

State s_2 State s_1 State s_0

The states s_0 and s_1 are absorbing: Once the play reaches one of these states it never leaves it. State s_2 is non-absorbing. Stochastic games with a single non-absorbing state are called *absorbing games*. For every $v = (v_0, v_1, v_2) \in \mathcal{V} = \mathbf{R}^3$ the game $G_{s_2}^\lambda(v)$ is the following matrix game:

$$\begin{array}{c|c|c|} & L & R \\ \hline T & (1-\lambda)v_2 & \lambda + (1-\lambda)v_1 \\ \hline B & \lambda + (1-\lambda)v_1 & (1-\lambda)v_0 \\ \hline \end{array}$$

The game $G_{s_2}^\lambda$

$$\begin{array}{c|c|} & L \\ \hline T & \lambda + (1-\lambda)v_1 \\ \hline \end{array} \qquad \begin{array}{c|c|} & L \\ \hline T & (1-\lambda)v_0 \\ \hline \end{array} \; .$$

The game $G_{s_1}^\lambda$ The game $G_{s_0}^\lambda$

The unique fixed point of the operator $\mathrm{val}(G^\lambda)$ must satisfy

- $\widehat{v}_0 = \mathrm{val}(G_{s_0}^\lambda(\widehat{v}))$, so that $\widehat{v}_{s_0}^\lambda = \widehat{v}_0 = 0$;
- $\widehat{v}_1 = \mathrm{val}(G_{s_1}^\lambda(\widehat{v}))$, so that $\widehat{v}_{s_1}^\lambda = \widehat{v}_1 = 1$;
- $\widehat{v}_2 = \mathrm{val}(G_{s_1}^\lambda(\widehat{v}))$. By Theorem 8 both players have a stationary λ-discounted 0-optimal strategy. Denote by x (resp. y) a mixed action for player 1 (resp. player 2) that is part of a λ-discounted 0-optimal strategy at the state s_2. Since we know that in the fixed point $\widehat{v}_0 = 0$ and $\widehat{v}_1 = 1$, $\widehat{v}_2$ must be the unique solution of

  $$v_2 = y(1-\lambda)v_2 + (1-y) = y \, ,$$

  so that $\widehat{v}_{s_2}^\lambda = \widehat{v}_2 = (1 - \sqrt{\lambda})/(1 - \lambda)$. The 0-optimal strategy of player 2 at state s_2 is $y = \widehat{v}_2 = (1 - \sqrt{\lambda})/(1 - \lambda)$, and the 0-optimal strategy of player 1, $x = \widehat{v}_2 = (1 - \sqrt{\lambda})/(1 - \lambda)$, can be found by finding his 0-optimal strategy in $G_{s_2}^\lambda(\widehat{v})$.

Bewley and Kohlberg [11] proved that when the state and action spaces are finite, the function $\lambda \mapsto v_s^\lambda$, that assigns to every state s and every discount factor λ the λ-discounted value at the initial state s, is a *Puiseux function*, that is, it has a representation $v_s^\lambda = \sum_{k=K}^\infty a_k \lambda^{k/M}$ that is valid for every $\lambda \in (0, \lambda_0)$ for some $\lambda_0 > 0$, where M is a natural number, K is a non-negative integer, and $(a_k)_{k=K}^\infty$ are real numbers. In particular, the function $\lambda \mapsto v_s^\lambda$ is monotone in a neighborhood of 0, and its limit as λ goes to 0 exists. This result turned out to be crucial in subsequent study on games with finitely many states and actions.

Shapley's work has been extended to general state and action spaces; for a recent survey see [46]. The tools developed in [46], together with a dynamic programming argument, prove that under proper conditions on the payoff function and on the transitions the two-player zero-sum stochastic game has a T-stage value.

Maitra and Sudderth [35] proved that the limsup value exists in a very general setup. Their proof follows closely that of Martin [36] for the determinacy of Blackwell games.

The study of the uniform value emanated from an example, called the "Big Match", due to Gillette [28], that was solved by Blackwell and Ferguson [13].

Example 10 Consider the following stochastic game with two absorbing states and one non-absorbing state.

$$\begin{array}{c|c|c|} & L & R \\ \hline T & 0 \;\; s^2 & 1 \;\; s^2 \\ \hline B & 1 \;\; s^1 & 0 \;\; s^0 \\ \hline \end{array} \qquad \begin{array}{c|c|} & L \\ \hline T & 1 \;\; s^1 \\ \hline \end{array} \qquad \begin{array}{c|c|} & L \\ \hline T & 0 \;\; s^0 \\ \hline \end{array}$$

State s_2 State s_1 State s_0

Suppose the initial state is s_2. As long as player 1 plays T the play remains at s_2; once he plays B the play moves to either s_0 or s_1, and is effectively terminated. By finding the fixed point of the operator φ one can show that the discounted value at the initial state s_2 is $\frac{1}{2}$, and a λ-discounted stationary 0-optimal strategy for player 2 is[1] $[\frac{1}{2}(L), \frac{1}{2}(R)]$. Indeed, if player 1 plays T then the expected stage payoff is $\frac{1}{2}$ and play remains at s_2, while if player 1 plays B then the game moves to an absorbing state, and the expected stage payoff from that stage onwards is $\frac{1}{2}$. In particular, this strategy guarantees $\frac{1}{2}$ for player 2 both in the limsup evaluation and uniformly. A λ-discounted 0-optimal strategy for player 1 is $[\frac{1}{1+\lambda}(T), \frac{\lambda}{1+\lambda}(B)]$.

What can player 1 guarantee in the limsup evaluation and uniformly? If player 1 plays the stationary strategy $[x(T), (1-x)(B)]$ that plays at each stage the action T with probability x and the action B with probability $1-x$, then player 2 has a reply that ensures that the limsup payoff is 0: If $x = 1$ and player 2 always plays L, the payoff is 0 at each stage; if $x < 1$ and player 2 always plays R, the payoff is 1 until the play moves to s_0, and then it is 0 forever. Since player 1 plays the action B with probability $1-x > 0$ at each stage, the distribution of the stage in which play moves to s_0 is geometric. Therefore, the limsup payoff is 0, and if λ is sufficiently small, the discounted payoff is close to 0.

One can verify that if player 1 uses a bounded-recall strategy, that is, a strategy that uses only the last k actions that were played, player 2 has a reply that guarantees that the limsup payoff is 0, and the discounted payoff is close to 0, provided λ is close to 0. Thus, in the limsup payoff and uniformly finite memory cannot guarantee more than 0 in this game (see also [27]).

Intuitively, player 1 would like to condition the probability of playing T on the past behavior of player 2: If in the past player 2 played the action L more often than the action R, he would have liked to play T with higher probability; if in the past player 2 played the action R more often than the action L, he would have liked to play B with higher probability. Blackwell and Ferguson [13] constructed a family of good strategies $\{\sigma_1^M, M \in \mathbf{N}\}$ for player 1. The parameter M determines the amount that the strategy guarantees: The strategy σ_1^M guarantees a limsup payoff and a discounted payoff of $\frac{M}{2M+1}$, provided the discount factor is sufficiently low. In other words, player 1 cannot guarantee $\frac{1}{2}$, but he may guarantee an amount as close to $\frac{1}{2}$ as he wishes by choosing M to be sufficiently large. The strategy σ_1^M is defined as follows: At stage t, play B with probability $\frac{1}{(M+l_t-r_t)^2}$, where l_t is the number of stages up to stage t in which player 2 played L, and r_t is the number of stages up to stage t in which player 2 played R.

Since $r_t + l_t = t-1$ one has $r_t - l_t = 2r_t - (t-1)$. The quantity r_t is the total payoff that player 1 received in the first $t-1$ stages if player 1 played T in those stages (and the game was not absorbed). Thus, this total payoff is a linear function of the difference $r_t - l_t$. When presented this way, the strategy σ_1^M depends on that total payoff. Observe that as r_t increases, $r_t - l_t$ increases as well, and the probability to play B decreases.

Mertens and Neyman [38] generalized the idea presented at the end of Example 10 to stochastic games with finite state and action spaces.[2]

Theorem 11 *If the state and action spaces of a two-player zero-sum stochastic game are finite, the game has a uniform value v_s^0 at every initial state $s \in S$. Moreover, $v_s^0 = \lim_{\lambda\to 0} v_s^\lambda = \lim_{T\to\infty} v_s^T$.*

In their proof, Mertens and Neyman describe a uniform ε-optimal strategy. In this strategy the player keeps a parameter, λ_t, which is a fictitious discount factor to use at stage t. This parameter changes at each stage as a function of the stage payoff; if the stage payoff at stage t is high then $\lambda_{t+1} < \lambda_t$, whereas if the stage payoff at stage t is low then $\lambda_{t+1} > \lambda_t$. The intuition is as follows. As mentioned before, in stochastic games there are two forces that influence the player's behavior: He tries to get high stage payoffs, while keeping future prospects high (by playing in such a way that the next stage that is reached is favorable). When considering the λ-discounted payoff there is a clear comparison between the importance of the two forces: The weight of the stage payoff is λ and the weight of future prospects is $1-\lambda$; the lower the discount factor, the more weight is given to the future. When considering the uniform value (or the uniform equilibrium) the weight of the stage payoff is 0. However, if the player never attempts to receive a high stage payoff, the overall payoff in the game will not be high. Therefore, the player has a fictitious discount factor; if past payoffs are low and they do not meet the expectation, player 1 increases the weight of the stage payoff by increasing the fictitious discount factor; if past payoffs are high, player 1 increases the weight of the future by lowering this fictitious discount factor.

Multi-Player Games

Takahashi [75] and Fink [21] extended Shapley's [60] result to discounted equilibria in non-zero-sum games.

[1] That is, at each stage player 2 plays L with probability $\frac{1}{2}$ and R with probability $\frac{1}{2}$.

[2] Mertens and Neyman's [38] result actually holds in every stochastic game that satisfies a proper condition, which is always satisfied when the state and action spaces are finite.

Theorem 12 *Every stochastic game with finite state and action spaces has a λ-discounted equilibrium in stationary strategies.*

Proof The proof utilizes Kakutani's fixed point theorem [31]. Let $M = \max_{i,s,a} |u_i(s, a)|$ be a bound on the absolute values of the payoffs. Set $X = \times_{i \in N, s \in S} (\Delta(A_i(s)) \times [-M, M])$. A point $x = (x^A_{i,s}, x^V_{i,s})_{i \in N, s \in S} \in X$ is a collection of one mixed action and one payoff to each player at every state. For every $v = (v_i)_{i \in N} \in [-M, M]^{N \times S}$ and every $s \in S$ define a matrix game $G^\lambda_s(v)$ as follows:

- The action spaces of each player i is $A_i(s)$;
- The payoff to player i is

$$\lambda u_i(s, a) + (1 - \lambda) \sum_{s' \in S} q(s' \mid s, a) v_i(s') .$$

We define a set-valued function $\varphi \colon X \to X$ as follows.

- For every $i \in N$ and every $s \in S$, $\varphi^A_{i,s}$ is the set of all best responses of player i to the strategy vector $x_{-i,s} := (x_{j,s})_{j \neq i}$ in the game $G^\lambda_s(v)$. That is,

$$\varphi^A_{i,s}(x, v) := \Big\{ \operatorname{argmax}_{y_{i,s} \in \Delta(A_i(s))} \lambda r_i(s, y_{i,s}, x_{-i,s}) + (1 - \lambda) \sum_{s' \in S} q(s' \mid s, y_{i,s}, x_{-i,s}) v_{i,s'} \Big\} .$$

- For every $i \in N$ and every $s \in S$, $\varphi^V_{i,s}(x, v)$ is the maximal payoff for player i in the game $G^\lambda_s(v)$, when the other players play x_{-i}:

$$\varphi^V_s(x, v) := \max_{y_{i,s} \in \Delta(A_i(s))} \Bigg(\lambda r(s, y_{i,s}, x_{-i,s}) + (1 - \lambda) \times \sum_{s' \in S} q(s' \mid s, y_{i,s}, x_{-i,s}) v_{i,s'} \Bigg) .$$

The set-valued function φ has convex and non-empty values and its graph is closed, so that by Kakutani's fixed point theorem it has a fixed point. It turns out that every fixed point of φ defines a λ-discounted equilibrium in stationary strategies. □

This result has been extended to general state and action spaces by various authors. These results assume a strong continuity on either the payoff function or the transition function, see, e. g., [39,44,62].

As in the case of zero-sum games, a dynamic programming argument shows that under a strong continuity assumption on the payoff function or on the transitions a T-stage equilibrium exists.

Regarding the existence of the limsup equilibrium and uniform equilibrium little is known. The most significant result in this direction is Vieille [77,78], who proved that every two-player stochastic game with finite state and action spaces has a uniform ε-equilibrium for every $\varepsilon > 0$. This result has been proved for other classes of stochastic games, see, e. g., [6,24,25,61,63,65,76]. Several influential works in this area are [22,32,71,79]. Most of the papers mentioned above rely on the vanishing discount factor approach, which constructs a uniform ε-equilibrium by studying a sequence of λ-discounted equilibria as the discount factor goes to 0.

For games with general state and action spaces, a limsup equilibrium exists under an ergodicity assumption on the transitions, see e. g. Nowak [44], Remark 4 and Jaskiewicz and Nowak [30].

A game has perfect information if there are no simultaneous moves, and both players observe past play. Existence of equilibrium in this case was proven by Mertens [37] in a very general setup.

Correlated Equilibrium

The notion of correlated equilibrium was introduced by Aumann [8,9], see also Forges ► Correlated Equilibria and Communication in Games. A *correlated equilibrium* is an equilibrium of an extended game, in which each player receives at the outset of the game a private signal such that the vector of signals is chosen according to a known joint probability distribution. In repeated interactions, such as in stochastic games, there are two natural notions of correlated equilibria: (a) each player receives one signal at the outset of the game (*normal-form correlated equilibrium*); (b) each player receives a signal at each stage (*extensive-form correlated equilibrium*). It follows from Forges [26] that when the state and action sets are finite, the set of all correlated T-stage equilibrium payoffs (either normal-form or extensive-form) is a polytope.

Nowak and Raghavan [47] proved the existence of an extensive-form correlated discounted equilibrium under weak conditions on the state and action spaces. In their construction, the strategies of the players are stationary, and so is the distribution according to which the signals are chosen after each history; both depend only on the current state, rather than on the whole past play. Roughly, their approach is to apply Kakutani's fixed point theorem to the set-valued function that assigns to each game $G^\lambda_s(v)$ the set of all correlated equilibrium payoffs in this game, which is convex and compact.

Solan and Vielle [66] proved the existence of an extensive-form correlated uniform equilibrium payoff when the state and action spaces are finite. Their approach is to let each player play his uniform optimal strategy in a zero-sum game in which all other players try to minimize his payoff. Existence of a normal-form correlated equilibrium was proved for the class of absorbing games [68].

Solan [64] characterized the set of extensive-form correlated equilibrium payoffs for general state and action spaces and a general evaluation on the stage payoffs, and provided a sufficient condition that ensures that the set of normal-form correlated equilibrium payoffs coincides with the set of extensive-form correlated equilibrium payoffs.

Imperfect Monitoring

So far it has been assumed that at each stage the players know the past play. There are cases in which this assumption is too strong; in some cases players do not know the complete description of the current state (Examples 3 and 4), and in others players do not fully observe the actions of all other players (Examples 2, 3 and 4). For a most general description of stochastic games see Mertens, see Chapter IV in Sorin and Zamir [40] and Coulomb [17].

In the study of the discounted equilibrium, the T-stage equilibrium or the limsup equilibrium, one may consider the game as a one-shot game: The players simultaneously choose strategies, and the payoff is either the discounted payoff, the T-stage payoff or the limsup payoff. If the strategy spaces of the players are compact (e. g., if the state and action spaces are finite), and if the payoff is upper-semi-continuous in each player's strategy, keeping the strategies of the other players fixed, then an equilibrium exists. This approach can be used successfully for the discounted equilibrium or the T-stage equilibrium under weak conditions (see, e. g., [4]), and may be used for the limsup equilibrium under a proper ergodicity condition.

Whenever there exists an equilibrium in stationary strategies (e. g., a discounted equilibrium in games with finitely many states and actions) the only information that players need in order to follow the equilibrium strategies is the current state. In particular, they need not observe past actions of the other players. As we now show, in the "Big Match" (Example 10) the limsup value and the uniform value may fail to exist when each player does not observe the past actions of the other player.

Example 13 (Example 10: Continued.) Assume that no player observes the actions of the other player, and assume that the initial state is s_2. Player 2 can still guarantee $\frac{1}{2}$ in the limsup evaluation by playing the stationary strategy $[\frac{1}{2}(L), \frac{1}{2}(R)]$. One can show that for every strategy of player 2, player 1 has a reply such that the limsup payoff is at least $\frac{1}{2}$. In other words, $\inf_{\sigma_2} \sup_{\sigma_1} \gamma^{\infty}(s_2, \sigma_1, \sigma_2) = \frac{1}{2}$. We now argue that $\sup_{\sigma_1} \inf_{\sigma_2} \gamma^{\infty}(s_2, \sigma_1, \sigma_2) = 0$. Indeed, fix a strategy σ_1 for player 1, and $\varepsilon > 0$. Let θ be sufficiently large such that the probability that under σ_1 player 1 plays B for the first time after stage t is at most ε. Observe that as t increases, the probability that player 1 plays B for the first time after stage t decreases to 0, so that such a θ exists. Consider the following strategy σ_2 of player 2: Play R up to stage θ, and play L from stage $t + 1$ and on. By the definition of θ, either player 1 plays B before or at stage θ, and then the game moves to s_0, and the payoff is 0 at each stage thereafter, or player 1 plays T at each stage, and then the stage payoff after stage θ is 0, or, with probability less than ε, player 1 plays B for the first time after stage θ, the play moves to s_1, and the payoff is 1 thereafter. Thus, the limsup payoff is at most ε. A similar analysis shows that

$$\inf_{\sigma_2} \sup_{\sigma_1} \gamma^{\lambda}(s_2, \sigma_1, \sigma_2) = \tfrac{1}{2} \,,$$

$$\sup_{\sigma_1} \inf_{\sigma_2} \lim_{\lambda \to 0} \gamma^{\lambda}(s_2, \sigma_1, \sigma_2) = 0 \,,$$

so that the uniform value does not exist as well.

This example shows that in general the limsup value and the uniform value need not exist when the players do not observe past play. Though in general the value (and therefore also an equilibrium) need not exist, in many classes of stochastic games the value and an equilibrium do exist, even in the presence of imperfect monitoring.

Rosenberg et al. [55] and Renault [54] showed that the uniform value exists in the one player setup (Markov Decision Problem), in which the player receives partial information regarding the current state. Thus, a single decision maker who faces a dynamic situation and does not fully observe the state of the environment can play in such a way that guarantees high payoff, provided the interaction is sufficiently long or the discount factor is sufficiently low.

Altman et al. [5,6] and Flesch et al. [24] studied stochastic games in which each player has a "private" state, which only he can observe, and the state of the world is composed of the vector of private states. Altman et al. [5,6] studied the situation in which players do not observe the actions of the other players, and Flesch et al. [24] studied the situation in which players do observe each others payoffs. Such games arise naturally in wireless communication (see [5]); take for example several mobiles who periodically send information to a base station. The private state of a mobile may depend, e. g., on its exact physical environment, and it determines the power attenuation between the mobile and the base station. The throughput (the amount of bits per second) that a mobile can send to the base station depends on the power attenuations of all the mobiles. Finally, the stage payoff is the stage power consumption.

Rosenberg et al. [57] studied the extreme case of two player zero-sum games in which the players observe neither the current state nor the action of the other player, and proved that the uniform value does exist in two classes of games, which capture the behavior of certain communication protocols. Classes of games in which the actions are observed but the state is not observed were studied, e. g., by Sorin [69,70], Sorin and Zamir [74], Krausz and Rieder [33], Flesch et al. [23], Rosenberg et al. [56], Renault [52,53]. For additional results, see [16,73].

Algorithms

There are two kinds of algorithms: Those that terminate in a finite number of steps, and those that iterate and approximate solutions. Both kinds of algorithms were devised to calculate the value and optimal strategies (or equilibria) in stochastic games.

It is well known that the value of a two-player zero-sum matrix game and optimal strategies for the two players can be calculated efficiently using a linear program. Equilibria in two-player non-zero-sum games can be calculated by the Lemke–Howson algorithm, which is usually efficient, however, its worst running time is exponential in the number of pure strategies of the players [59]. Unfortunately, to date there are no efficient algorithms to calculate either the value in zero-sum stochastic games, or equilibria in non-zero-sum games. Moreover, in Example 9 the discounted value may be irrational for rational discount factors, even though the data of the game (payoffs and transitions) are rational, so it is not clear whether linear programming methods can be used to calculate the value of a stochastic game. Nevertheless, linear programming methods were used to calculate the discounted and uniform value of several classes of stochastic games, see [20,50]. Other methods that were used to calculate the value or equilibria in discounted stochastic games include fictitious play [80], value iterates, policy improvement, and general methods to find the maximum of a function (see [20,51]), a homotopy method [29], and algorithms to solve sentences in formal logic [15,67].

Additional and Future Directions

The research on stochastic games extends to additional directions than those mentioned in earlier sections. We mention a few here. Approximation of games with infinite state and action spaces by finite games was discussed by Whitt [81], and further developed by Nowak [43]. Stochastic games in continuous time have also been studied, as well as hybrid models that include both discrete and continuous aspects, see, e. g., [2,10].

Among the many directions of future research in this area, we will mention here but a few. One challenging question is the existence of a uniform equilibrium and a limsup equilibrium in multi-player stochastic games with finite state and action spaces. Another is the development of efficient algorithms that calculate the value of two-player zero-sum games. A third direction concerns the identification of applications that can be recast in the framework of stochastic games, and that can be successfully analyzed using the theoretical tools that the literature developed. Another problem that is of interest is the characterization of approachable and excludable sets in stochastic games with vector payoffs (see [12] for the presentation of matrix games with vector payoffs, and [41] for partial results regarding this problem).

Acknowledgments

I thank Eitan Altman, János Flesch, Yuval Heller, Jean-Jacques Herings, Ayala Mashiach-Yakovi, Andrzej Nowak, Ronald Peeters, T.E.S. Raghavan, Jérôme Renault, Nahum Shimkin, Robert Simon, Sylvain Sorin, William Sudderth, and Frank Thuijsman, for their comments on an earlier version of the entry.

Bibliography

Primary Literature

1. Altman E (2005) Applications of dynamic games in queues. Adv Dyn Games 7:309–342
2. Altman E, Gaitsgory VA (1995) A hybrid (differential-stochastic) zero-sum game with fast stochastic part. Ann Int Soc Dyn Games 3:47–59
4. Altman E, Solan E (2007) Games with constraints with networking applications. Preprint
5. Altman E, Avrachenkov K, Marquez R, Miller G (2005) Zero-sum constrained stochastic games with independent state processes. Math Methods Oper Res 62:375–386
6. Altman E, Avrachenkov K, Bonneau N, Debbah M, El-Azouzi R, Sadoc Menasche D (2008) Constrained cost-coupled stochastic games with independent state processes. Oper Res Lett 36:160–164
7. Amir R (1996) Continuous stochastic games of capital accumulation with convex transitions. Games Econ Behav 15:111–131
8. Aumann RJ (1974) Subjectivity and correlation in randomized strategies. J Math Econ 1:67–96
9. Aumann RJ (1987) Correlated equilibrium as an expression of bayesian rationality. Econometrica 55:1–18
10. Başar T, Olsder GJ (1995) Dynamic noncooperative game theory. Academic Press, New York
11. Bewley T, Kohlberg E (1976) The asymptotic theory of stochastic games. Math Oper Res 1:197–208

12. Blackwell D (1956) An analog of the minimax theorem for vector payoffs. Pac J Math 6:1–8
13. Blackwell D, Ferguson TS (1968) The big match. Ann Math Stat 39:159–163
14. Chari V, Kehoe P (1990) Sustainable plans. J Political Econ 98:783–802
15. Chatterjee K, Majumdar R, Henzinger TA (2008) Stochastic limit-average games are in EXPTIME. Int J Game Theory 37:219–234
16. Coulomb JM (2003) Absorbing games with a signalling structure. In: Neyman A, Sorin S (eds) Stochastic games and applications. NATO Science Series. Kluwer, Dordrecht, pp 335–355
17. Coulomb JM (2003) Games with a recursive structure. In: Neyman A, Sorin S (eds) Stochastic games and applications. NATO Science Series. Kluwer, Dordrecht, pp 427–442
18. Dutta P, Sundaram RK (1992) Markovian equilibrium in a class of stochastic games: Existence theorems for discounted and undiscounted models. Econ Theory 2:197–214
19. Dutta P, Sundaram RK (1993) The tragedy of the commons? Econ Theory 3:413–426
20. Filar JA, Vrieze K (1996) Competitive Markov decision processes. Springer
21. Fink AM (1964) Equilibrium in a stochastic *n*-person game. J Sci Hiroshima Univ 28:89–93
22. Flesch J, Thuijsman F, Vrieze K (1997) Cyclic Markov equilibria in stochastic games. Int J Game Th 26:303–314
23. Flesch J, Thuijsman F, Vrieze OJ (2003) Stochastic games with non-observable actions. Math Meth Oper Res 58:459–475
24. Flesch J, Schoenmakers G, Vrieze K (2008) Stochastic games on a product state space. Math Oper Res 33:403–420
25. Flesch J, Thuijsman F, Vrieze OJ (2007) Stochastic games with additive transitions. Europ J Oper Res 179:483–497
26. Forges F (1990) Universal mechanisms. Econometrica 58:1341–1364
27. Fortnow L, Kimmel P (1998) Beating a finite automaton in the big match. In: Proceedings of the 7th conference on theoretical aspects of rationality and knowledge. Morgan Kaufmann, San Francisco, pp 225–234
28. Gillette D (1957) Stochastic games with zero stop probabilities, contributions to the theory of games, vol 3. Princeton University Press, Princeton
29. Herings JJP, Peeters RJAP (2004) Stationary equilibria in stochastic games: Structure, selection, and computation. J Econ Theory 118:32–60
30. Jaskiewicz A, Nowak AS (2006) Zero-sum ergodic stochastic games with Feller transition probabilities. SIAM J Control Optim 45:773–789
31. Kakutani S (1941) A generalization of Brouwer's fixed point theorem. Duke Math J 8:457–459
32. Kohlberg E (1974) Repeated games with absorbing states. Ann Stat 2:724–738
33. Krausz A, Rieder U (1997) Markov games with incomplete information. Math Meth Oper Res 46:263–279
34. Levhari D, Mirman L (1980) The great fish war: An example using a dynamic Cournot–Nash solution. Bell J Econ 11(1):322–334
35. Maitra A, Sudderth W (1998) Finitely additive stochastic games with Borel measurable payoffs. Int J Game Theory 27:257–267
36. Martin DA (1998) The determinacy of Blackwell games. J Symb Logic 63:1565–1581
37. Mertens JF (1987) Repeated games. In: Proceedings of the international congress of mathematicians, American Mathematical Society, Berkeley, California, pp 1528–1577
38. Mertens JF, Neyman A (1981) Stochastic games. Int J Game Th 10:53–66
39. Mertens JF, Parthasarathy T (1987) Equilibria for discounted stochastic games, CORE Discussion Paper No. 8750. (Also published in Stochastic Games and Applications, Neyman A, Sorin S (eds), NATO Science Series, Kluwer, 131–172)
40. Mertens JF, Sorin S, Zamir S (1994) Repeated games, CORE Discussion Paper 9420-9422
41. Milman E (2006) Approachable sets of vector payoffs in stochastic games. Games Econ Behav 56:135–147
42. Neyman A, Sorin S (2003) Stochastic games and applications. NATO Science Series. Kluwer
43. Nowak AS (1985) Existence of equilibrium stationary strategies in discounted noncooperative stochastic games with uncountable state space. J Optim Theory Appl 45:591–620
44. Nowak AS (2003) *N*-person stochastic games: Extensions of the finite state space case and correlation. In: Neyman A, Sorin S (eds) Stochastic games and applications. NATO Science Series. Kluwer, Dordrecht, pp 93–106
45. Nowak AS (2003) On a new class of nonzero-sum discounted stochastic games having stationary Nash equilibrium points. Int J Game Theory 32:121–132
46. Nowak AS (2003) Zero-sum stochastic games with Borel state spaces. In: Neyman A, Sorin S (eds) Stochastic games and applications. NATO Science Series. Kluwer, Dordrecht, pp 77–91
47. Nowak AS, Raghavan TES (1991) Existence of stationary correlated equilibria with symmetric information for discounted stochastic games. Math Oper Res 17:519–526
48. Phelan C, Stacchetti E (2001) Sequential equilibria in a Ramsey tax model. Econometrica 69:1491–1518
49. Puterman ML (1994) Markov decision processes: Discrete stochastic dynamic programming. Wiley, Hoboken
50. Raghavan TES, Syed Z (2002) Computing stationary Nash equilibria of undiscounted single-controller stochastic games. Math Oper Res 27:384–400
51. Raghavan TES, Syed Z (2003) A policy improvement type algorithm for solving zero-sum two-person stochastic games of perfect information. Math Program Ser A 95:513–532
52. Renault J (2006) The value of Markov chain games with lack of information on one side. Math Oper Res 31:490–512
53. Renault J (2007) The value of repeated games with an informed controller. Preprint
54. Renault J (2007) Uniform value in dynamic programming. Preprint
55. Rosenberg D, Solan E, Vieille N (2002) Blackwell optimality in Markov decision processes with partial observation. Ann Statists 30:1178–1193
56. Rosenberg D, Solan E, Vieille N (2004) Stochastic games with a single controller and incomplete information. SIAM J Control Optim 43:86–110
57. Rosenberg D, Solan E, Vieille N (2006) Protocol with no acknowledgement. Oper Res, forthcoming
58. Sagduyu YE, Ephremides A (2003) Power control and rate adaptation as stochastic games for random access. Proc 42nd IEEE Conf Decis Control 4:4202–4207
59. Savani R, von Stengel B (2004) Exponentially many steps for finding a Nash equilibrium in a bimatrix game. Proc 45th Ann IEEE Symp Found Comput Sci 2004:258–267

60. Shapley LS (1953) Stochastic games. Proc Nat Acad Sci USA 39:1095–1100
61. Simon RS (2003) The structure of non-zero-sum stochastic games. Adv Appl Math 38:1–26
62. Solan E (1998) Discounted stochastic games. Math Oper Res 23:1010–1021
63. Solan E (1999) Three-person absorbing games. Math Oper Res 24:669–698
64. Solan E (2001) Characterization of correlated equilibria in stochastic games. Int J Game Theory 30:259–277
65. Solan E, Vieille N (2001) Quitting games. Math Oper Res 26:265–285
66. Solan E, Vieille N (2002) Correlated equilibrium in stochastic games. Games Econ Behav 38:362–399
67. Solan E, Vieille N (2007) Calculating uniform optimal strategies and equilibria in two-player stochastic games. Preprint
68. Solan E, Vohra R (2002) Correlated equilibrium payoffs and public signalling in absorbing games. Int J Game Theory 31:91–122
69. Sorin S (1984) Big match with lack of information on one side (part 1). Int J Game Theory 13:201–255
70. Sorin S (1985) Big match with lack of information on one side (part 2). Int J Game Theory 14:173–204
71. Sorin S (1986) Asymptotic properties of a non-zerosum stochastic games. Int J Game Theory 15:101–107
72. Sorin S (2002) A first course on zero-sum repeated games. Mathématiques et Applications, vol 37. Springer
73. Sorin S (2003) Stochastic gameswith incomplete information. In: Neyman A, Sorin S (eds) Stochastic Games and Applications. NATO Science Series. Kluwer, Berlin, pp 375–395
74. Sorin S, Zamir S (1991) Big match with lack of information on one side (part 3). In: Raghavan TES et al (eds) Stochastic games and related topics. Kluwer, pp 101–112
75. Takahashi M (1962) Stochastic games with infinitely many strategies. J Sci Hiroshima Univ Ser A-I 26:123–134
76. Thuijsman F, Raghavan TES (1997) Perfect information stochastic games and related classes. Int J Game Theory 26:403–408
77. Vieille N (2000) Equilibrium in 2-person stochastic games I: A Reduction. Israel J Math 119:55–91
78. Vieille N (2000) Equilibrium in 2-person stochastic games II: The case of recursive games. Israel J Math 119:93–126
79. Vrieze OJ, Thuijsman F (1989) On equilibria in repeated games with absorbing states. Int J Game Theory 18:293–310
80. Vrieze OJ, Tijs SH (1982) Fictitious play applied to sequences of games and discounted stochastic games. Int J Game Theory 12:71–85
81. Whitt W (1980) Representation and approximation of noncooperative sequential games. SIAM J Control Optim 18:33–48

Books and Reviews

Başar T, Olsder GJ (1995) Dynamic noncooperative game theory. Academic

Filar JA, Vrieze K (1996) Competitive Markov decision processes. Springer

Maitra AP, Sudderth WD (1996) Discrete gambling and stochastic games. Springer

Mertens JF (2002) Stochastic games. In: Aumann RJ, Hart S (eds) Handbook of game theory with economic applications, vol 3. Elsevier, pp 1809–1832

Mertens JF, Sorin S, Zamir S (1994) Repeated games. CORE Discussion Paper 9420-9422

Raghavan TES, Shapley LS (1991) Stochastic games and related topics: In honor of Professor L.S. Shapley. Springer

Vieille N (2002) Stochastic games: Recent results. In: Aumann RJ, Hart S (eds) Handbook of game theory with economic applications, vol 3. Elsevier, pp 1833–1850

Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems

Hans C. Fogedby[1,2]

[1] Department of Physics and Astronomy, University of Aarhus, Aarhus, Denmark

[2] Niels Bohr Institute, Copenhagen, Denmark

Article Outline

Glossary

Bessel process The Bessel process operates in d dimensions and describes the radial distance r from the origin of a particle performing Brownian motion. The random motion is governed by the Langevin equation $\mathrm{d}r/\mathrm{d}t = \kappa(d-1)/2r + \xi$, where $\langle\xi\xi\rangle(t) = \kappa\delta(t)$. For $d \leq 2$ the motion is recurrent, i. e., returns to the origin; for $d > 2$ the motion goes off to infinity. In SLE the Bessel process describes the transition between simple curves and self-intersecting curves.

Brownian motion Brownian motion, B_t, is the scaling limit of random walk. Brownian motion is plane-filling, has the fractal dimension $D = 2$, and is described by the Langevin equation $\mathrm{d}B_t/\mathrm{d}t = \eta_t, \langle\eta_t\eta_s\rangle = \delta(t-s)$. B_t is characterized by i) the stationarity property, $B_{t+t'} - B_t$ and $B_{t'}$ identical in distribution and ii) the independence property, $B_{\Delta t'}$ and $B_{\Delta t}$ independent for $\Delta t \neq \Delta t'$. The correlations are given by

$\langle |B_t - B_s|^2 \rangle = |t - s|$ and B_t is distributed according to the Gaussian (normal) distribution $P(B, t) = (2\pi t)^{-1/2} \exp(-B^2/2t)$.

Chordal SLE In order to map the geometry of the growing hull to a real function by means of SLE one chooses a reference domain. Chordal SLE refers to the case where the SLE trace is grown between two boundary points, usually the origin and the point at infinity in the upper half complex plane.

Conformal invariance Conformal invariance or local scale invariance is a larger symmetry than scale invariance. Conformal invariance implies invariance under both a local rotation, translation, and dilatation. Conformal invariance is particularly powerful in 2D where a conformal transformation is implemented by an analytic function.

Conformal transformation A conformal transformation is a transformation which preserves angles but allows for rotation, dilatation, and translation. In an elastic medium picture a conformal transformation corresponds to translation, rotation, and compression (dilatation), but not shear. In 2D an analytic complex function $w = f(z)$ from the complex z plane to the complex w plane generates a conformal transformation.

Continuum limit The limit where a lattice model approaches a continuum model; also called the scaling limit. The resulting continuum field theories define the universality classes of the lattice models.

Correlation length The correlation length measures the size of correlations. At the critical point the correlation length diverges signaling that the system becomes scale invariant.

Critical curves Critical curves are domain walls or cluster boundaries at the critical point. Critical curves are scale invariant and characterized by a fractal dimension.

Exploration Domain walls at the critical point can mathematically be generated by an exploration process where the domain wall is initiated at a boundary point and constructed in steps across the domain. In the percolation case the local step is generated by 'flipping a coin', in the Ising case by evaluating the magnetization at the tip of the 'growing' domain wall. The construction by an exploration process is essential in generating a critical curve by stochastic Loewner evolution.

Fortuin–Kasteleyn representation (FK) Based on a high temperature expansion one can represent the configurations in the Ising, $O(n)$, and Potts models by means of the Fortuin–Kasteleyn representation in terms of random clusters. The FK transformation is essential in identifying random curves for the lattice models which then can be accessed by SLE in the scaling limit.

Fractal dimension Irregular objects can be characterized by a fractal dimension. The fractal dimension D is derived by covering the object with $N(\ell)$ intervals, disks, or spheres of linear dimension ℓ. By letting $\ell \to 0$ this definition resolves the fine scale fractal structure and the scaling relation $N(\ell) \propto (\ell)^{-D}$, or $D = -\lim_{\ell \to 0} \ln(N(\ell)/\ln(\ell)$, yields the fractal dimension. Examples of deterministic fractals are for example the Cantor set, the Koch curve, the Sierpinski gasket, and plane-filling Hilbert and Peano curves. Random walk and diffusion limited aggregation (DLA) are examples of random physical fractals.

Hulls For a self-intersecting SLE trace the trace and enclosed regions form a so-called hull. Points in the hull cannot be connected to infinity without crossing the trace. The growing hull of a self-intersecting SLE trace eventually exhausts the half plane.

Ising model The Ising model originates from the theory of magnetism and plays an important role in the theory of phase transitions and in general in statistical mechanics. The model is defined on a lattice where each lattice sites is endowed with a local spin variable assuming two distinct values. The spins interact by a short range exchange interaction. Above 1D the Ising model has a second order phase transition.

Locality In the percolation case the domain wall is constructed by an exploration process and a local rule for assigning the next step. This locality property is specific to percolation which has a geometric phase transition. In the SLE context the locality property implies $\kappa = 6$, i. e., the percolation case.

Loewner equation The Loewner equation is the first order nonlinear differential equation for the uniformizing map $g_t(z)$ which maps the complement of a growing hull or curve in the upper half plane back to the upper half plane. The geometrical properties of the hull is encoded in the real function a_t. The Loewner equation has the form $\mathrm{d}g_t/\mathrm{d}t = 2/(g_t - a_t)$.

Loewner evolution Loewner evolution refers to the parametrized map $g_t(z)$ which uniformizes the complement of a curve or hull, say in the upper half plane.

Loop erased random walk (LERW) Loop erased random walk is a variant of random walk where loops formed are removed as the walk progresses. LERW has a conformally invariant scaling limit and can be generated by SLE for $\kappa = 2$.

Markov property The Markov property refers to a stochastic process without memory. A typical example is random walk or in the scaling limit Brownian motion. The Markov property is essential for the application of SLE to domain walls in the scaling limit.

Measures For lattice systems one can define probability distributions according to the rules of statistical mechanics. In the scaling limit the distributions typically diverge and have to be replaced by the more abstract mathematical concept of probability measures.

$O(n)$ models The $O(n)$ model is a generalization of the Ising model. At each lattice site is associated an n-component spin variable. For $n = 1$ we recover the Ising model, $n = 2$ corresponds to the XY model, and $n = 3$ to the Heisenberg model.

Peano curve A Peano curve is a non-crossing curve which is dense in the plane, i. e., it gets arbitrarily close to every point. The Peano curve has the fractal dimension $D = 2$. In a SLE context a random Peano curve winds around a UST and corresponds to $\kappa = 8$.

Percolation Percolation on a lattice is constructed by filling lattice sites at random with a common probability p. At a critical concentration a spanning cluster extends across the system.

Potts model The Potts model is a generalization of the Ising model where the local lattice variable can assume q different values and where sites only interact when they are in the same Potts state. The Potts model has a FK cluster representation. The Ising model corresponds to $q = 2$; $q = 1$ corresponds to percolation and $q = 0$ to the UST.

Random walk Random walk is an ubiquitous phenomenon in nature. In random walk on a square lattice a particle jump from site to site with a given fixed probability. Each step is independent of the past history, there is no memory, this is the so-called Markov property. The random walk path or history is plane-filling and has the fractal dimension $D = 2$. The mean square displacement of random walk scales with the number of steps.

Restriction The restriction property in a SLE context implies that the measure on a curve in a domain D conditioned not to hit a bulge L is the same as the measure in the domain $D \setminus L$. The restriction property holds in the case of a uniform measure and applies to self-avoiding random walk.

Riemann's mapping theorem Riemann's mapping theorem states that an arbitrary simply connected domain D, i. e., without holes, can be mapped to another simply connected domain D' by means of a suitable complex function $g(z)$, i. e., $g(D) = D'$. Often the disk or half plane are used as reference domains. The mapping theorem does not make assumptions about the domain boundaries which can be fractal.

Scale invariance The scaling property refers to the case where a phenomenon is devoid of a characteristic scale or unit. Scale invariance is typically characterized by a power law behavior and critical exponents.

Scaling limit The limit where the lattice parameter in a lattice model approaches zero. The scaling limit is equivalent to the continuum limit.

Self-avoiding random walk (SAW) A self-avoiding random walk is a walk conditioned not to cross itself. SAW has been used to model polymers. SAW has a uniform probability measure and conforms in a SLE context to the restriction condition.

Stochastic Loewner evolution (SLE) Stochastic Loewner evolution is Loewner evolution driven by a real stochastic function a_t with a distribution given by 1D Brownian motion, i. e., $a_t = \sqrt{\kappa} B_t$. SLE is governed by the stochastic equation of motion $\mathrm{d}g_t/\mathrm{d}t = 2/(g_t - a_t)$.

Schramm's theorem Schramm's theorem refers to Schramm's derivation of SLE for LERW. Schramm showed that the scaling limit of LERW is described by SLE for $\kappa = 2$. Schramm also conjectured that the random Peano curve winding around an UST is described by SLE for $\kappa = 8$ and that percolation is described by SLE for $\kappa = 6$.

Uniformizing maps Conformal transformations which map a domain D to a standard reference domain, e. g., the half plane or the disk, are called uniformizing maps.

Uniform spanning tree (UST) A spanning tree is a collection of vertices and links forming a tree (no loops or cycles). A USF is a random tree picked among all spanning trees with equal probability.

Definition of the Subject

Stochastic Loewner evolution also called Schramm Loewner evolution (abbreviated, SLE) is a rigorous tool in mathematics and statistical physics for generating and studying scale invariant or fractal random curves in two dimensions (2D). The method is based on the older deterministic Loewner evolution introduced by Karl Löwner [64], who demonstrated that an arbitrary curve not crossing itself can be generated by a real function by means of a conformal transformation. A real function defined in one spatial dimension (1D) thus encodes a curve in 2D, in itself an intriguing result. In 2000 Oded Schramm [77] extended this method and demonstrated that driving the Loewner evolution by a 1D Brownian motion, the curves in the complex plane become scale invariant; the fractal dimension turns out to be determined by the strength of the Brownian motion.

The one parameter family of scale invariant curves generated by SLE is conjectured and has in some cases been proven to represent the continuum or scaling limit of a variety of interfaces and cluster boundaries in lattice models in statistical physics, ranging from self-avoiding random walks to percolation cluster boundaries, and Ising domain walls.

SLE operates in the 2D continuum where it generates extended scale invariant objects. SLE delimits scaling universality classes by a single parameter κ, the strength of the 1D Brownian drive, yielding the fractal dimension D of the scale invariant shapes according to the relation $D = 1 + \kappa/8$. Moreover, SLE provides the geometrical aspects of conformal field theory (CFT). The central charge c, delimiting scaling universality classes in CFT, is thus related to κ by means of the expression $c = (3\kappa - 8)(6 - \kappa)/2\kappa$.

Stochastic Loewner evolution derives its importance from the fact that it addresses the issue of extended random fractal shapes in 2D by direct analysis in the continuum. It thus supplements and extends earlier lattice results and also allows for the determination of new scaling exponents. From the point of view of conformal field theory based on the concept of a local field, operator expansions, and correlations, the geometrical approach afforded by SLE, directly addressing conformally invariant random shapes in the continuum, represents a novel point of view of maybe far reaching consequences; so far only explored in two dimensions.

Introduction

The field of SLE has mainly been driven by mathematicians presenting their results in long and difficult papers. There are, however, presently several excellent reviews of SLE both addressing the theoretical physics community [11,13,25,27,39,43] and the mathematical community [56,85], see also a complete biography up to 2003 [39]. The purpose of this article is to present a heuristic and simple discussion of some of the key aspects of SLE, for details and topics left out we refer the reader to the reviews mentioned above. However, in order to provide the necessary background and set the stage for SLE we begin with some general remarks on scaling in statistical physics.

General Remarks

In statistical physics we study macroscopic systems composed of many interacting components. In the limit of many degrees of freedom the macroscopic behavior roughly falls in two categories. In the most common case the macroscopic behavior is deterministic and governed by phenomenological theories like for example thermodynamics and hydrodynamics operating entirely on a macroscopic level. This behavior is basically a result of the law of large numbers, permitting an effective coarse graining and yielding for example a macroscopic density or velocity field [29]. In the other case, the macroscopic behavior is dominated by fluctuations and shows a random behavior [29,73]. Typical cases are random walk and equilibrium systems tuned to the critical point of a second order phase transition. Other random cases are for example self-organized critical systems purported to model earth quake dynamics, flicker noise and turbulence in fluids [5,40].

The distinction between the deterministic and random cases of macroscopic behavior is illustrated by the simple example of a biased random walk described by the Langevin equation $\mathrm{d}x(t)/\mathrm{d}t = v + \xi(t)$, $\langle\xi(t)\xi(0)\rangle \propto \delta(t)$. Here $x(t)$ is a macroscopic variable sampling the statistically independent microscopic steps $\xi(t)$; the velocity v is an imposed drift or bias. Averaging over the steps we obtain for the deviation of x, $R = [\langle x^2\rangle]^{1/2} = [(vt)^2 + t]^{1/2}$. For large t the deviation $R \sim \langle x\rangle = vt$ and the mean value or deterministic part dominates the behavior, the fluctuational or random part $R \sim t^{1/2}$ being subdominant. Fine tuning the random walk to vanishing bias $v = 0$ we have $\langle x\rangle = 0$ and $R \sim t^{1/2}$, i. e., the random fluctuations control the phenomenon.

The study of complexity encompasses a broader field than statistical physics and is concerned with the emergence of universal properties on a mesoscopic or macroscopic scale in large interacting systems. For example particle systems, networks in biology and sociology, cellular automata, etc. The class of complex systems generally falls in the category of random systems. The emergent properties are a result of many interacting agents or degrees of freedom and can in general not be directly inferred from the microscopic details. A major issue is thus the understanding of generic emerging properties of complex systems [45,70,83]. Here, however, the methods of statistical physics is an indispensable tool in the study of complexity.

The evolution of statistical physics, a branch of theoretical physics, has occurred in steps and is driven both by the introduction of new concepts and the concurrent development of new mathematical methods and techniques. A well-known case is the long standing problem of second order phase transitions or critical phenomena which gave way to a deeper understanding in the sixties and seventies and spun off the renormalization group techniques for the determination of critical exponents and universality classes [20,24,65,81,86].

Scaling Ideas

This brings us to the fundamental scaling ideas and techniques developed particularly in the context of critical phenomena in equilibrium systems and which now pervade a good part of theoretical physics and, moreover, play an important role in the analysis of complex systems in physical sciences [20,24,29]. Scaling is synonymous with no scale in the sense that a system exhibiting scale invariance is characterized by the absence of any particular scale or unit. Scaling occurs both in the space and/or time behavior and is typically characterized by power law dependencies controlled by scaling exponents.

A classical case is random walk discussed above, characterized by the Langevin equation $\mathrm{d}x/\mathrm{d}t = \xi(t)$, $\langle\xi\xi\rangle \sim \delta(t)$ [73]. Here the mean square displacement scales like $\langle x^2\rangle(t) \sim t^{2H}$, where H is the Hurst scaling exponent; for random walk $H = 1/2$ [32,68]. Correspondingly, the power spectrum $P(\omega) = |x_\omega|^2$, $x_\omega = \int \mathrm{d}t x(t) \exp(i\omega t)$, scales like $P(\omega) \sim \omega^{-1-2H}$, i. e., $P(\omega) \sim \omega^{-2}$ for random walk; we note that the underlying reason for the universal scaling behavior of random walk is the central limit theorem [37,73].

Scaling in Equilibrium

Scaling ideas and associated techniques first came to the forefront in statistical physics in the context of second order phase transitions or critical phenomena [20,24,29,65]. More specifically, consider the usual Ising model defined on a lattice with a local spin degrees of freedom, $\sigma_i = \pm 1$ at site i, subject to a short range interaction J favoring spin alignment. The Ising Hamiltonian has the form $H = -J\sum_{\langle ij\rangle}\sigma_i\sigma_j$, where $\langle ij\rangle$ indicates nearest neighbor sites. The thermodynamic phases are characterized by the order parameter $m = \langle\sigma_i\rangle$. Above one dimension the Ising model exhibits a second order phase transition at a finite critical temperature T_c. Above T_c the model is in a disordered paramagnetic state with $m = 0$ and only microscopic domain of ordered spins. Below T_c the model favors a ferromagnetic state with long range order and macroscopic domains of ordered spins, corresponding to $m \neq 0$; at $T = 0$ the model assumes the ferromagnetic ground state configuration with totally aligned spins. Regarding the spatial organization, the size of the domains of ordered spins is characterized by the correlation length $\xi(T)$. As we approach the critical point at T_c the order parameter vanishes $m \propto |T - T_c|^\beta$ with critical exponent β, but more significantly, the correlation length $\xi(T)$ diverges like $\xi(T) \sim |T - T_c|^{-\nu}$ with scaling exponent ν. This indicates that the system is scale invariant at T_c. Regarding the domains of ordered spins, the system is spatially self-similar on scales from the microscopic lattice distance to the system size; the system size diverging in the thermodynamic limit. The scaling behavior at the critical point T_c is an emergent property in the sense that the scaling exponents β and ν do not depend on microscopic details like the type of lattice, strength of interaction, etc., but only on the dimension of the system and the symmetry of the order parameter [72].

The diverging correlation length at the critical point is the central observation which in the 60-ties and 70-ties gave rise to a detailed understanding of critical phenomena, beginning with the coarse graining block scheme proposed by Kadanoff [41] and culminating with the development and application of field theoretical renormalization group techniques by Wilson and others [20,24,29,65,72,86].

For a diverging correlation length much larger than the lattice distance the local spin σ_i can be replaced by a coarse-grained local field $\phi(r)$ and the Ising Hamiltonian H by the Ginzburg–Landau functional $F = \int \mathrm{d}^d r[(\nabla\phi)^2 + R\phi^2 + U\phi^4]$, where the 'mass' term $R \sim |T - T_c|$. Consequently, the universality class of Ising-type models is described by a scalar field theory. The renormalization group techniques basically quantify the Kadanoff block construction in momentum space and extract scaling properties in an expansion about the upper critical dimension $d = 4$. To leading order in $4 - d$ one obtains $\beta = 1/2 - (4 - d)/6$ and $\nu = 1/2 + (4 - d)/12$. Alternatively, keeping the correlation length ξ fixed and letting the lattice distance approach zero, we obtain at T_c the so-called scaling limit or continuum limit of the Ising model. The Ising spin σ_i becomes a local field $\phi(r)$ and the weight of a configuration is determined by $\exp(-F)$. Note that in order to implement the scaling limit we must be at T_c. The two scenarios of a growing correlation length for fixed lattice distance implementing the Kadanoff construction and a fixed correlation length for a vanishing lattice distance yielding a continuum field theory are related by an overall scale transformation [86].

It is generally assumed that the global or nonlocal scale invariance at the critical point in the continuum limit can be extended to a local scale invariance including translation and rotation, that is an angle-preserving conformal transformation. This follows heuristically from an a local implementation of the Kadanoff coarse-graining block construction and applies to lattice models with short range interactions and discrete translational and rotational invariance [22,24]. The resulting continuum theories then fall in the category of conformal field theories (CFT).

In 2D the group of conformal transformations is particularly rich since it corresponds to the class of analytical functions $w = f(z)$, mapping the complex plane z to the complex plane w. The infinite group structure imposes sufficient constraints on the structure of conformal field theories in 2D that the scaling form of correlations, e. g., $\langle\phi\phi\rangle(r)$, and in particular the critical exponents can be determined explicitly [23,24]. On finds $\beta = 1/8$ and $\nu = 1$ for the order parameter and correlation length exponents, respectively, in accordance with lattice theory results [14]. Here we also mention the Coulomb gas method for the determination of critical exponents [71].

It is a common feature of both renormalization group calculations based on an expansion about a critical dimension and conformal field theory in 2D that the local field $\phi(r)$ and its correlations are the basic building blocks and that the critical properties are encoded in their scaling form, yielding critical exponents, scaling laws, scaling functions, etc. Notwithstanding the fact that the seminal Kadanoff construction [41] was based on a geometrical picture corresponding to coarse-graining the degrees of freedom over larger and larger scales, keeping track of ordered domains on all scales, the actual geometry of critical phenomena such as the scaling properties of critical clusters was not well-understood and seemed inaccessible in the continuum limit within the context of local field theories. Whereas it is not difficult to generate critical domain walls, interfaces, and clusters for lattice models with appropriate boundary conditions by means of standard Monte Carlo simulation techniques, the continuum or scaling limit of critical shapes appeared until recently, with a few exceptions, beyond present techniques.

Stochastic Loewner Evolution

Here stochastic Loewner evolution (SLE) represents a new insight in 2D critical phenomena with respect to a deeper understanding of scale invariant curves, clusters, and shapes. Also, there appears to be deep connections between SLE and conformal field theory.

SLE is an ingenious way of generating critical curves and shapes in the 2D continuum using conformal transformations. Let us mention a characteristic example. Consider an Ising model on a lattice in a chosen domain. Imposing spin up on a continuous part of the domain boundary and spin down on the remaining part of the boundary, it follows that a specific domain wall or interface will connect the two points on the boundary where a bond is broken. At low temperature the bond energies dominate and the free energy is lowest for a straight domain wall with few kinks. As we approach the critical point entropy or fluctuations come strongly into play and the domain wall meanders balancing energy and entropy. At the critical point the system becomes scale invariant with a diverging correlation length and likewise the domain wall becomes scale invariant, i. e., it has kinks on all scales larger that the lattice distance. In the continuum limit the Ising domain wall becomes a random fractal curve with a particular fractal dimension. Here SLE provides a direct analytical method in the continuum to generate such a random curve and, moreover, provides the fractal dimension in terms of the strength of the 1D random walk driving the SLE evolution.

Outline

The outline of the present article is as follows. In Sect. "Scaling" on scaling we introduce some of the basic models and concepts necessary for a discussion of SLE: A) Random walk, B) Percolation, C) Ising model, D) Critical curves and exploration, and E) Distributions, Markov properties, and measures. In Sect. "Conformal Invariance" we turn to the essential ingredient in SLE, namely, conformal invariance: A) Conformal maps and B) Measures and conformal invariance. Section "Loewner Evolution" is devoted to deterministic or classical Loewner evolution: A) Growing stick, B) Loewner equation and C) Exact solutions. In Sect. "Stochastic Loewner Evolution", constituting the core part of this article, we discuss stochastic Loewner evolution: A) Schramm's theorem, B) SLE properties, C) Curves, hulls, and the Bessel process and D) Fractal dimension. Section "Results and Discussion" is devoted to results and discussions: A) Phase transitions, locality, restriction, and duality, B) Loop erased random walk, C) Self-avoiding random walk, D) Percolation, E) Ising model and $O(n)$ models, F) SLE and conformal field theory, G) Application to 2D turbulence, H) Application to 2D spin glass and I) Further remarks. Finally, in Sect. "Future Directions" we discuss future directions of the field. The bibliography includes books, general reviews, and more technical papers.

Scaling

Stochastic Loewner evolution, has been applied to a host of lattice models proved or conjectured to possess scaling limits. However, for the present purpose we will focus on three lattice models: Random walk, Percolation, and the Ising model.

Random Walk

Random walk is a simple and much studied random process [4,32,73]. Consider an unbiased random walk in the plane composed of N steps, where the ith step, $\vec{\eta}_i$, is random, isotropic and uncorrelated, i. e., $\langle\vec{\eta}_i\rangle = 0$ and $\langle\eta_i^\alpha\eta_j^\beta\rangle \propto \delta_{\alpha\beta}\delta_{ij}$. For the end-to-end distance we have $\vec{x} = \sum_{i=1}^N \vec{\eta}_i$ and for the size $R = [\langle\vec{x}^2\rangle]^{1/2} \sim N^{1/2}$. Assuming one step pr unit time, $N \propto t$, we obtain $R \sim t^{1/2}$, characteristic of diffusive motion. Introducing the fractal dimension D by the usual box counting procedure [32,68]

$$N(R) \sim R^D\,, \tag{1}$$

where N is the number of boxes and R the size of the object, and covering the random walk we readily infer $D = 2$, i. e., the self-crossing random walk is plane-filling modulo the lattice distance. Introducing the scaling exponent ν according to $R \sim N^\nu$ we have for random walk $\nu = 1/2$; note that $D = 1/\nu$.

The scaling limit of unbiased random walk is Brownian motion (BM) [4] and is obtained by scaling the step size η down and the number of steps N up in such a manner that the size $R \sim N^{1/2}\eta$ stays constant. The resulting BM path is a continuous non-differentiable random curve with fractal dimension $D = 2$. The BM path is plane-filling and recurrent in 2D, i. e., it returns to a given point with probability one. Focusing on one of the independent cartesian components 1D BM, B_t, is also described by the Langevin equation

$$\frac{\mathrm{d}B_t}{\mathrm{d}t} = \eta_t\,, \quad \langle\eta_t\eta_s\rangle = \delta(t-s)\,, \tag{2}$$

where η_t is uncorrelated Gaussian white noise with a flat power spectrum. Integrating Eq. (2) B_t samples the step η_t and we find, assuming $B_0 = 0$, $B_t = \int_0^t \eta_{t'}\mathrm{d}t'$ from which we directly infer the fundamental properties of BM, namely, independence and stationarity,

$$B_{T+\Delta T} - B_T \approx B_{\Delta T}\,, \qquad \text{(stationarity)} \tag{3}$$

$$B_{\Delta T}, B_{\Delta T'} \ \text{ indep. for } \ \Delta T \neq \Delta T'\,, \quad \text{(independence)} \tag{4}$$

where $\approx$ indicates identical distributions. Moreover, the correlations are given by

$$\langle|B_t - B_s|^2\rangle = |t-s| \tag{5}$$

and B_t is distributed according to the normal (Gaussian) distribution $P(B,t) = (2\pi t)^{-1/2}\exp(-B^2/2t)$.

Whereas 1D BM drives SLE, 2D BM is not itself generated by SLE since the path is self-crossing on all scales; by construction SLE is limited to the generation of non-crossing random curves. However, variations of BM have played an important role in the development of SLE. We mention here the scaling limit of loop erased random walk (LERW) and self-avoiding random walk (SAW), to be discussed later.

Percolation

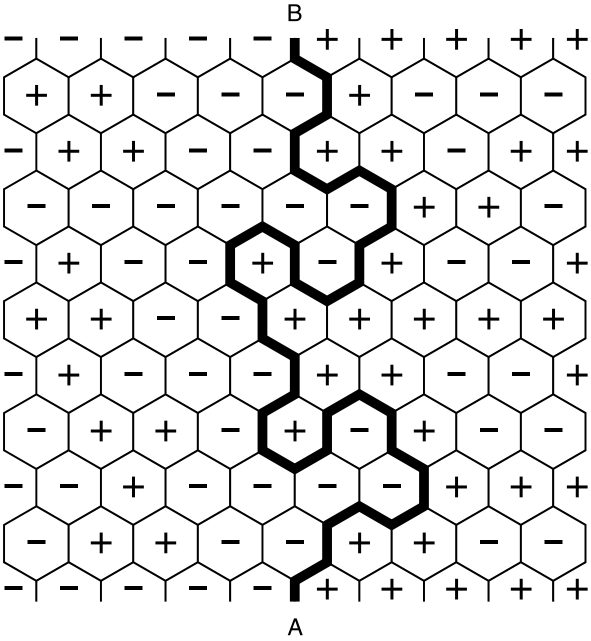
The phenomenon of percolation is relevant in the context of clustering, diffusion, fractals, phase transitions and disordered systems. As a result, percolation theory provides a theoretical and statistical background to many physical and natural sciences [82].

Percolation is the simplest lattice model exhibiting a geometrical phase transition. The site percolation model is constructed by occupying sites on a lattice with a given common probability p. Let an occupied site be denoted 'plus' and an 'empty' site denoted 'minus'. For p close to zero the sites are mainly unoccupied and the lattice is 'minus'. For p close to one the sites are predominantly occupied and the lattice is 'plus'. At a critical concentration p_c, the percolation threshold, an infinite cluster of 'plus' sites embedded in the 'minus' background extends across the lattice. In the scaling limit of vanishing lattice distance the critical cluster has a fractal boundary which can be accessed by SLE. Whereas the scaling properties of critical percolation clusters define a universality class and is independent of the lattice structure, the critical concentration p_c in general depends on the lattice. For site percolation on a triangular lattice the percolation threshold is known to be $p_c = 1/2$.

In Fig. 1 we depict a realization of site percolation on a triangular lattice in the upper half plane at the percolation threshold $p_c = 1/2$. The occupied sites are denoted 'plus', the empty sites 'minus'. By imposing appropriate boundary conditions we induce a meandering domain wall across the system from A to B. In the scaling limit the domain wall becomes a fractal non-crossing critical curve.

Ising Model

The Ising model is probably the simplest interacting many particle system in statistical physics [20,29,81]. The model has its origin in magnetism but has become of paradigmatic importance in the context of phase transitions. The model is defined on a lattice where each lattice site i is occupied by a single degree of freedom, a spin variable σ_i, assuming two values, $\sigma_i = \pm 1$, i. e., spin up or spin down.



Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems, Figure 1
We depict site percolation on a triangular lattice in the *upped half plane* at the percolation threshold. The critical concentration is $p_c = 1/2$. The occupied sites are denoted 'plus', the empty sites 'minus'. The boundary conditions enforce a meandering domain wall from *A* to *B*

In the ferromagnetic case considered here the spins interact via a short range exchange interaction J favoring parallel spin alignment. The model is described by the Ising Hamiltonian

$$H = -J \sum_{\langle ij \rangle} \sigma_i \sigma_j \,, \tag{6}$$

where $\langle ij \rangle$ indicates nearest neighbor spin sites i and j. The statistical weight or probability of a specific spin configuration $\{\sigma_i\}$ is determined by the Boltzmann factor

$$P(\{\sigma_i\}) = \exp[-H/kT]/Z \,, \tag{7}$$

where T is the temperature and k Boltzmann's constant. The partition function Z has the form

$$Z = \sum_{\{\sigma_i\}} \exp(-H/kT) \,, \tag{8}$$

yielding the thermodynamic free energy F according to $F = -kT \log Z$. The entropy is given by $S = -\mathrm{d}F/\mathrm{d}T$ and the energy follows from $F = E - TS$. The magnetization or order parameter and correlations are given by

$$m = \sum_{\{\sigma_i\}} \sigma_i P(\{\sigma_i\}) \quad \text{and} \quad \langle \sigma_i \sigma_j \rangle = \sum_{\{\sigma_i\}} \sigma_i \sigma_j P(\{\sigma_i\}) \,, \tag{9}$$

respectively.

The Ising model possesses a phase transition at a critical temperature T_c from a disordered paramagnetic phase above T_c with vanishing order parameter $m = 0$ to a ferromagnetic ordered phase below T_c with non-vanishing order parameter $m \neq 0$. At the critical temperature T_c the order parameter vanishes like $m \sim |T - T_c|^{\beta}$ with critical exponent β. The correlation function $\langle \sigma_i \sigma_j \rangle$ monitoring the spatial organization of ordered domains behaves like

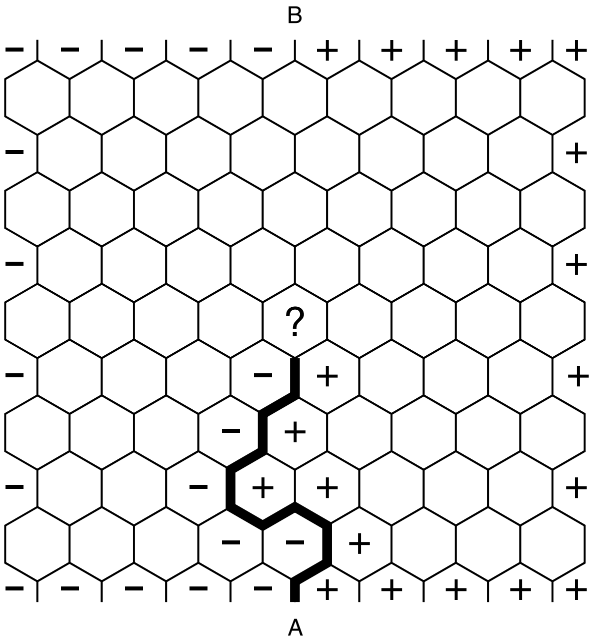
$$\langle \sigma_i \sigma_j \rangle \sim \frac{\exp[-|i-j|/\xi]}{|i-j|^{\eta}} \,, \tag{10}$$

where $|i - j|$ denotes the distance between position i and position j. The algebraic behavior is characterized by the critical exponent η and the correlation length ξ scales like $\xi \sim |T - T_c|^{-\nu}$ with critical exponent ν. At the critical point the correlation length ξ diverges and the Ising model becomes scale invariant with an algebraically decaying correlation function $\langle \sigma_i \sigma_j \rangle \sim |i - j|^{-\eta}$. Assigning 'plus' to spin up and 'minus' to spin down, Fig. 1 also illustrates a typical configuration of the 2D Ising model on a triangular lattice at the critical temperature, including an Ising domain wall from A to B.

Critical Curves – Exploration

In the percolation case at the percolation threshold a critical curve is induced by fixing the boundary conditions. Occupying sites from A to B along the right hand side of the boundary and assigning empty sites along the left hand side of the boundary from A to B a critical curve will meander across the system from A to B. Imagining painting the two sides of the curve, one side is painted with occupied sites, the other side with empty sites. Typically the curve meanders on all scales but by construction does not cross itself. For later purposes the configuration is depicted in Fig. 1.

In order to make contact with SLE we observe that a critical interface in the percolation case also can be constructed by an exploration process. We initiate the curve at the boundary point A and toss a coin. In the case of 'head' the site or hexagon in front is chosen to be occupied and the path bends left; in the event of 'tail' the hexagon in front is left unoccupied and the path bends right. In this manner a critical meandering non-crossing curve is generated terminating eventually at B. The percolation growth



Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems, Figure 2

We depict the growth process in the percolation case. The percolation threshold is at $p_c = 1/2$. The interface imposed by the boundary conditions originates at the boundary point *A* and progresses towards the boundary point *B*

process is depicted in Fig. 2, where we for clarity only have indicated the sites involved in the growth process. We note that since there is no interaction between the sites the path depends entirely on the local properties.

In the case of the Ising model a critical interface or domain walls at the critical point is again fixed by assigning appropriate boundary conditions with spin up along the boundary from A to B and spin down from B to A. The critical curve can be constructed in two ways: Globally or by an exploration process. In the global case we generate a spin configuration by means of a Monte Carlo simulation, i. e., perform a biased importance sampling implementing the probability distribution in Eq. (7), and identify a critical interface. Alternatively, we can generate an interface by an exploration process like in the percolation case, occupying a site i according to the weight $(1/2)(1 + \langle\sigma_i\rangle)$, where $\langle\sigma_i\rangle$ is evaluated in the domain with the spins along the interface fixed, see again Fig. 2.

Distributions – Markov Properties – Measures

In the case of random walk the number of walks of length L grows like μ^L, where μ is a lattice dependent number, i. e., at each step there are μ lattice-dependent choices for choosing a direction of the next step. Consequently, the weight or probability of a particular walk of length L is proportional to μ^{-L},

$$P(L) \sim \mu^{-L} \,, \tag{11}$$

and all walks of a given length have the same weight. We note here the important Markov property, characteristic of random walk, which can be formulated in the following manner. Assume that the first part γ' of the walk has taken place and thus conditions the remaining part γ of the walk. In a suggestive notation the conditional distribution is given by $P(\gamma|\gamma') = P(\gamma\gamma')/P(\gamma')$. The Markov property implies that the conditional probability is equal to the probability of the walk γ in a new domain where the first part of the walk γ' has been removed, i. e., the identity

$$P(\gamma|\gamma') = P'(\gamma) \,, \quad \text{(Markov property)} \tag{12}$$

where the prime refers to the new domain. The Markov property is self-evident for random walk and follows directly from Eq. (11), i. e., $P(\gamma|\gamma') = P(\gamma\gamma')/P(\gamma') = \mu^{-(L+L')}/\mu^{-L'} = \mu^{-L} = P'(\gamma)$, where L and L' are the lengths of segments γ' and γ, respectively.

In the scaling limit the lattice distance goes to zero whereas the length of the walk diverges. Consequently, the distribution diverges and must be replaced by an appropriate probability measure [4,13,43]. However, in order to define the probability distribution or measure in the scaling limit we shall assume that the Markov property continues to hold and interpret the P in Eq. (12) as probability measures. The Markov property is essential in carrying over the lattice probability distributions in the scaling limit.

For the critical interfaces defined by an exploration processes in the case of site percolation, the Markov property follows by inspection since the propagation of the interface is entirely determined by the local process of occupying the next site with probability 1/2. This locality property is specific for percolation which has a geometrical phase transition and does in the SLE context, to be discussed later, determine the scaling universality class.

In the case of an interface in the Ising model the Markov property also holds, but since the spins interact a little calculation is required [13]. Consider an interface γ defined by an exploration process. According to the rules of statistical mechanics the probability distribution for the interface γ is given by

$$P(\gamma) = \frac{Z(\gamma)}{Z} \,. \tag{13}$$

Here Z is the full partition function defined in Eq. (8) with appropriate boundary conditions imposed. $Z(\gamma)$ is the partial partition function with the spins associated with the interface γ fixed,

$$Z(\gamma) = \sum_{\{\sigma_i\},\gamma} \exp[-H(\gamma)/kT] \,. \tag{14}$$

The Hamiltonian $H(\gamma)$ inferred from Eq. (6) is the energy of a spin configuration with the spins determining γ fixed; $\{\sigma_i\},\gamma$ indicate the configurations to be summed over. Evidently, we have the identity $Z = \sum_\gamma Z(\gamma)$, i. e., $\sum_\gamma P(\gamma) = 1$.

Whereas in the random walk case we only considered an individual path and in the percolation case the interface only feels the nearby sites, an Ising interface is imbedded in the interacting spin systems and we have to define the Markov property more precisely with respect to a domain D. Consider an interface across the domain D from A to B and assume that the last part γ is conditioned on the determination of the first part γ', i. e., given by the distribution $P_D(\gamma|\gamma')$. Next imagine that we cut the domain D along the interface γ', i. e., break the interaction bonds between the spins determining γ'. The right and left face of γ' can then be considered part of the domain boundary and the Markov property states that the distribution of γ in the cut domain $D \setminus \gamma'$ (i. e., D minus γ') equals $P_D(\gamma|\gamma')$,

$$P_D(\gamma|\gamma') = P_{D\setminus\gamma'}(\gamma) \,. \quad \text{(Markov property)} \tag{15}$$

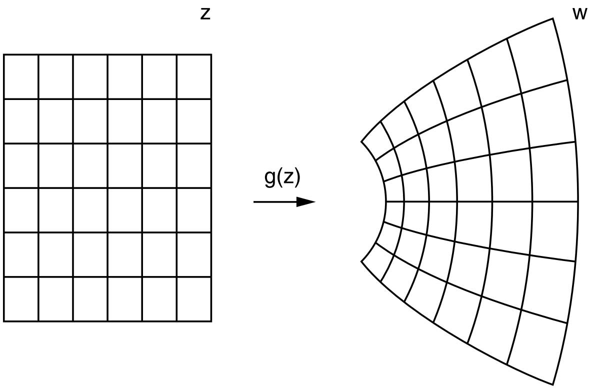
In order to demonstrate Eq. (15) we use the definition in Eq. (13). The conditional probability $P_D(\gamma|\gamma') = P_D(\gamma\gamma')/P_D(\gamma') = (Z_D(\gamma\gamma')/Z_D)/(Z_D(\gamma')/Z_D) = Z_D(\gamma\gamma')/Z_D(\gamma')$. Correspondingly, the distribution in the cut domain $D \setminus \gamma'$ is $P_{D\setminus\gamma'}(\gamma) = Z_{D\setminus\gamma'}(\gamma)/Z_{D\setminus\gamma'}$. However, it follows from the structure of the partition function in Eqs. (6–8) that $Z_{D\setminus\gamma'} = \exp[E(\gamma')/kT]Z_D(\gamma')$ and $Z_{D\setminus\gamma'}(\gamma) = \exp[E(\gamma')/kT]Z_D(\gamma\gamma')$, where $E(\gamma')$ is the energy of the broken bonds. By insertion the interface Boltzmann factors cancel out and we obtain Eq. (15) expressing the Markov property.

Conformal Invariance

In the complex plane analysis implies geometry. The representation of a complex number $z = x + iy$ directly associates complex function theory with 2D geometrical shapes. This connection is of importance in mathematical physics in for example 2D electrostatics and 2D hydrodynamics. In the context of SLE Riemann's mapping theorem plays an essential role [1].

Conformal Maps

Briefly, Riemann's mapping theorem [2] states that any simply connected domain, i. e., topologically deformable to a disk, in the complex plane z can be uniquely mapped to a unit disk $|w| < 1$ in the complex w plane by mean of a complex function $w = g(z)$. By combining complex functions we can map any simply connected domain to any other simply connected domain. For example, if $g_1(z)$ and $g_2(z)$ map domains D_1 and D_2 to the unit disk, respectively, then $g_2^{-1}(g_1(z))$ maps D_1 to D_2; here g_2^{-1} is the inverse function of g_2, i. e., $g_2^{-1}(g_2(z)) = z$. As an example, the transformation $g(z) = i(1 + z)/(1 - z)$ maps the the unit disk centered at the origin to the upper half plane. Likewise, the Möbius transformation $g(z) = (az + b)/(cz + d)$ determined by four real parameters, $ad - bc > 0$, maps the upper half plane onto itself. Conformal transformations are angle-preserving and basically correspond to a combination of a local rotation, local translation, and local dilatation. In terms of an elastic medium picture conformal transformations are equivalent to deformations without shear [54]. Expressing $g(z)$ in terms of its real and imaginary parts, $g(z) = u(x, y) + iv(x, y)$, the Cauchy–Riemann equations $\partial u/\partial x = \partial v/\partial y$ and $\partial u/\partial y = -\partial v/\partial x$ hold implying that u and v are harmonic functions satisfying Laplace's equations $\nabla^2 u = 0$ and $\nabla^2 v = 0$, $\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$. In Fig. 3 we have depicted a conformal angle-preserving transformation effectuated by the complex function $g(z)$ from the complex z plane to the complex w plane.



Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems, Figure 3
We depict a conformal transformation from the complex z plane to the complex w plane. We note the angle-preserving property, i. e., a shear-free transformation. The map in the figure is given by $w = z^2$

Measures – Conformal Invariance

Whereas the Markov property discussed above holds for lattice curves even away from criticality, we here want to assume another property which only holds in the scaling limit at the critical point, namely conformal invariance. In the scaling limit we anticipate that the probability measure $P(\gamma)$ for an interface γ is invariant under a conformal transformation. More precisely, consider a lattice model, say the Ising or percolation model, and specify two domains D and D' on the lattice. Next, consider an interface, cluster boundary or domain wall γ from the boundary points A and B across the domain D. In terms of the partition functions the probability distribution for γ is given by Eq. (13). We now perform the scaling or continuum limit of the lattice model keeping the domains D and D' fixed. The continuous random interface approaches its scaling form and is characterized by the measure $P_D(\gamma)$. At the critical point we assume that the interface is scale invariant under the larger symmetry of conformal transformations. The next step is to consider a specific conformal transformation $g(z)$ which according to Riemann's theorem precisely maps domain D to domain D', i. e., $D' = g(D)$ and the interface to $g(\gamma)$. The assumption of conformal invariance then states that the probability measure P is invariant under this transformation expressing the scale invariance, i. e.,

$$P_D(\gamma) = P_{g(D)}(g(\gamma)) \, . \quad \text{(Conformal property)} \qquad (16)$$

Both the Markov property and the conformal property are sufficient in combination with Loewner evolution to determine the measures in the scaling limit.

Loewner Evolution

The original motivation of Loewner's work was to examine the so called Bieberbach conjecture which states that $|a_n| \leq n$ for the coefficients in the Taylor expansion $f(z) = \sum_{n=0} a_n z^z$, where $f(z)$ maps the unit disk to the complex plane. The conjecture was proposed in 1916 and finally proven in 1984 by de Branges [2,38,39]. For that purpose Loewner [64] considered growing parametrized conformal maps to a standard domain. In the present context Loewner's method allow us to access growing shapes in 2D in an indirect manner by means of a 'time dependent' conformal transformation $g_t(z)$.

Growing Stick

Before we address the derivation of the Loewner equation let us consider the specific conformal transformation

$$g_t(z) = \sqrt{z^2 + 4t} \, . \qquad (17)$$

For $t = 0$ we have $g_0(z) = z$, i. e., the identity map. Likewise, for $z \to \infty$ we obtain

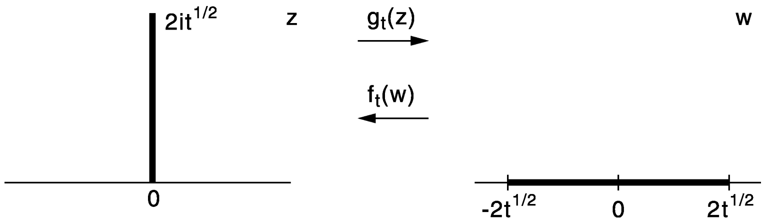
$$g_t(z) \sim z + \frac{2t}{z} \, , \qquad (18)$$

showing that far away in the complex plane we again have the identity map. The coefficient in the next leading term, C_t/z, is called the capacity C_t; here parametrized by the 'time variable', $t = C_t/2$. The map (17) has a branch point at $z = 2it^{1/2}$ and it follows by inspection that g_t maps the upper half plane minus a stick from the origin $\mathbf{0}$ to $2it^{1/2}$ back to the upper half plane. From the inverse map $f_t(w) = g_t^{-1}(w)$,

$$f_t(w) = \sqrt{w^2 - 4t} \, , \qquad (19)$$

we infer that the right face of the stick is mapped to the real axis from 0 to $2t^{1/2}$, the tip to the origin, and the left face to the interval $-2t^{1/2}$ to 0. Under the map the growing stick thus becomes part of the boundary in the w plane. The growing stick is depicted in Fig. 4.

The stick shows up as an imaginary contribution along the interval $-2t^{1/2}$ to $2t^{-1/2}$. More precisely, since $f_t(w)$ is analytic in the upper half plane, implementing the asymptotic behavior $f_t(w) \sim w$ for $w \to \infty$, and using Cauchy's



Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems, Figure 4

We depict the growing stick corresponding to the conformal transformation $w = \sqrt{z^2 + 4t}$. The vertical cut in the z plane extends from the origin to the point $(0, 2it^{1/2})$. The right and left faces of the cut are mapped to the real axis from $-2t^{1/2}$ to $+2t^{1/2}$, the endpoint to the origin, in the complex w plane

theorem, we obtain the dispersion relation or spectral representation

$$f_t(w) = w - \int \frac{\mathrm{d}\omega}{\pi} \frac{A_t(\omega)}{w-\omega} \,, \tag{20}$$

with spectral weight $A_t(\omega)$. In the case of the growing stick we find $A_t(\omega) = (4t - \omega^2)^{1/2}$ for $\omega^2 < 4t$ and otherwise $A_t(\omega) = 0$. Using $1/(\omega + i\epsilon) = \mathrm{P}\ 1/\omega - i\pi\delta(\omega)$ (P denotes principal value) we also have $\mathrm{Im} f_t(\omega) = A_t(\omega)$ and $\mathrm{Re} f_t(\omega) = \omega - \mathrm{P} \int (\mathrm{d}\omega'/\pi) A_t(\omega')/(\omega - \omega')$. The time dependent spectral weight $A_t(\omega)$ thus characterizes the growing stick. With the chosen parametrization we also have the sum rule

$$\int \frac{\mathrm{d}\omega}{\pi} A_t(\omega) = C_t = 2t \,. \tag{21}$$

Finally, we note that the map g_t satisfies the equation of motion

$$\frac{\mathrm{d}g_t(z)}{\mathrm{d}t} = \frac{2}{g_t(z)} \,, \tag{22}$$

i. e., solving Eq. (22) with the initial condition $g_0(z) = z$ and the boundary condition $g_t(z) \sim z$ for $z \to \infty$ we arrive at the map in Eq. (17).

Loewner Equation

The growing stick nicely illustrates the idea of accessing a growing shape indirectly by the application of Riemann's theorem mapping the domain adjacent to the shape to a standard reference domain, here the upper half plane. This so-called uniformizing map effectively absorbs the shape and encodes the information about the shape into the spectral weight $A_t(\omega)$ along the real axis.

Let us consider a general shape or hull $\mathbf{K_t}$ in the upper half plane $\mathbf{H}$. Together with the real axis the shape form part of the boundary of the domain $\mathbf{D}$. In other words, the domain in question is the upper half plane $\mathbf{H}$ with the shape $\mathbf{K_t}$ subtracted, $\mathbf{D} = \mathbf{H} \setminus \mathbf{K_t}$. Applying Riemann's theorem we map the simply connected domain $\mathbf{D}$ back to the upper half plane $\mathbf{H}$ by means of the conformal transformation $g_t(z)$, i. e., g_t absorbs the shape $\mathbf{K_t}$. Imagine that the shape grows a little bit further from $\mathbf{K_t}$ to $\mathbf{K}_{t+\delta t} = \mathbf{K_t} + \delta\mathbf{K_t}$, where $\mathbf{K_t}$ is now part of $\mathbf{K}_{t+\delta t}$; $\delta\mathbf{K_t}$ is the shape increment. Correspondingly, the map $g_{t+\delta t}$ is designed to absorb $\mathbf{K}_{t+\delta t}$, i. e., $\mathbf{H} \setminus \mathbf{K}_{t+\delta t} \to \mathbf{H}$ by means of the map $g_{t+\delta t}$. We now carry out the elimination in two ways. Either we absorb $\mathbf{K_t}$ by means of the map g_t and subsequently $\delta\mathbf{K_t}$ by means of the map δg_t or we absorb $\mathbf{K}_{t+\delta t}$ directly in one step by means of the map $g_{t+\delta t}$. Consequently, combining maps we have $g_{t+\delta t}(z) = \delta g_t(g_t(z))$ or $g_t(z) = \delta g_t^{-1}(g_{t+\delta t}(z))$. Since $\delta g_t^{-1}(w)$ is analytic in $\mathbf{H}$ we obtain the spectral representation

$$\delta g_t^{-1}(w) = w - \int \frac{\mathrm{d}\omega}{\pi} \frac{\delta A_t(\omega)}{w - \omega} \,, \tag{23}$$

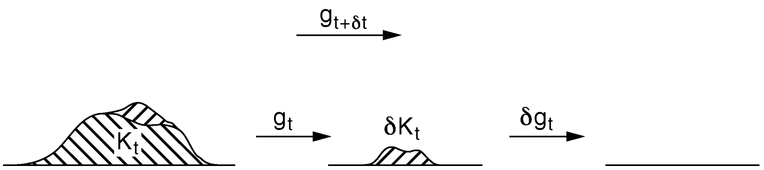
with infinitesimal spectral weight δA_t, or inserting $w = g_{t+\delta}(z)$

$$g_t(z) = g_{t+\delta t}(z) - \int \frac{\mathrm{d}\omega}{\pi} \frac{\delta A_t(\omega)}{g_{t+\delta t} - \omega} \,. \tag{24}$$

The last step is to set $\delta A_t(\omega) = \rho_t(\omega)\delta t$, yielding a differential equation for the evolution of the map g_t eliminating the shape $\mathbf{K_t}$,

$$\frac{\mathrm{d}g_t(z)}{\mathrm{d}t} = \int \frac{\mathrm{d}\omega}{\pi} \frac{\rho_t(\omega)}{g_t(z) - \omega} \,. \tag{25}$$

Specifying the weight or measure $\rho_t(\omega)$ along the real ω-axis this equation determines, through the uniformizing map g_t, how the shape $\mathbf{K_t}$ grows. The spectral weight encodes the 2D shape into the real function $\rho_t(\omega)$. Note that since $\rho_t(\omega)$ is not specified and can depend nonlinearly on the map, Eq. (25) still represent a highly nonlinear problem. Invoking the asymptotic condition $g_t(z) \sim z + C_t/z$, where C_t is the time dependent capacity we infer $\mathrm{d}C_t/\mathrm{d}t = \int (\mathrm{d}\omega/\pi)\rho_t(\omega)$ or since $\rho_t(\omega) = \mathrm{d}A_t(\omega)/\mathrm{d}t$ the sum rule in Eq. (21). The conformal mapping procedure is depicted in Fig. 5.



Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems, Figure 5

The combination of maps involved in the derivation of the Loewner equation. First the map g_t eliminates the hull K_t. Subject to the growth in the time interval δt the incremental hull δK_t is subsequently absorbed by the infinitesimal map δg_t. Correspondingly, the hull $K_{t+\delta t}$ is absorbed by the map $g_{t+\delta t}$ in one step

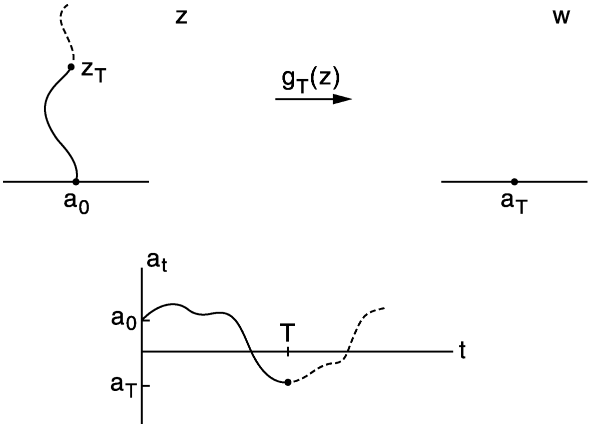
In the special case where the growth takes place at a point the equation simplifies considerably. Assuming that the spectral weight is concentrated at the point $\omega = a_t$, where a_t is a real function of t and setting $\rho_t(\omega) = 2\pi\delta(\omega - a_t)$ we arrive at the Loewner equation

$$\frac{\mathrm{d}g_t(z)}{\mathrm{d}t} = \frac{2}{g_t - a_t} . \tag{26}$$

The Loewner equation describes the growth of a curve or trace γ_t with endpoint z_t, $0 < t < \infty$, in the upper half complex z-plane. The time-dependent conformal transformation g_t maps the simply connected domain $\mathbf{H} \setminus \gamma_t$, i. e., the half plane excluding the curve γ_t back to the w half plane. At a given time instant t the tip of the curve z_t is determined by $g_t(z_t) = a_t$, i. e., the point where Eq. (26) develops a singularity. The topological properties and shape of the curve are encoded in the real function a_t which lives on the real axis in the w-plane. As a_t develops in time the tip of the curve z_t determined by $z_t = g_t^{-1}(a_t)$ traces out a curve. Since the domain $\mathbf{H} \setminus \gamma_\mathrm{t}$ must be simply connected for the Riemann theorem to apply the curve or trace cannot cross itself or cross the real axis. Whenever the curve touches or intersect itself or the real axis the enclosed part will be excluded from the domain. In other, words, during the time progression the curve effectively absorbs part of the upper half plane. It is a deep property of Loewner evolution that the topological properties of a 2D non-crossing curve are entirely encoded by the real function a_t. The encoding works both ways: A given 2D non-crossing curve γ_t corresponds to a specific real function a_t, a given real function a_t yields a specific 2D non-crossing curve γ_t. A continuous a_t will yield a continuous curve γ_t. A discontinuous a_t in general gives rise to branching. Whether or not the curve intersects or touches itself is determined by the singularity structure of the drive a_t. In the case where the Hölder condition $\lim_{\tau\to 0} |(a_{t+\tau} - a_\tau)/\tau^{1/2}|$ is greater that 4 we have self-intersection. Note again that since the curve is defined indirectly by the singularity structure in Eq. (26) we cannot easily identify a curve parametrization and for example determine a tangent vector, etc. The mechanism underlying the Loewner equation is shown in Fig. 6.

Exact Solutions

In a series of simple cases one can solve the Loewner equation analytically [42,44]. For vanishing drive $a_t = 0$ we obtain the growing vertical stick discussed above. Correspondingly, a constant drive $a_t = a$ yields a vertical stick growing up from the point a on the real axis.



Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems, Figure 6
The mechanism in the Loewner equation. The curve in the upper half complex z plane generated by the Loewner equation is mapped onto a finite but growing segment of the real axis of the complex w plane. The endpoint z_T is mapped to the real number a_T. As a_t develops in time and makes excursions along the real axis the endpoint z_t of the curve grows into the upper half plane

In the case of a linear drive $a_t = t$ the tip of the curve z_t is given by $z_t = 2 - 2\phi_t \cos\phi_t + 2i\phi_t$, where the phase ϕ_t is determined from the equations: $2\ln r_t - r_t \cos\phi_t = 2\ln 2 + t - 2$ and $r_t = 2\phi_t/\sin\phi_t$. By inspection $\phi_0 = 0$ and $\phi_\infty = \pi$. The curve thus approaches the asymptote $2\pi i$ for $t \to \infty$. For small t analysis yields $z_t \sim (2/3)t + 2i\sqrt{t}$, i. e., the trace approaches the origin with infinite slope. The square root drives $a_t = 2\sqrt{\kappa t}$ and $a_t = 2\sqrt{\kappa(1-t)}$, $0 < t < 1$ with a finite-time singularity can also be treated. In the first case, $a_t = 2\sqrt{\kappa t}$, the trace is a straight line $z_t = B\exp(i\phi)\sqrt{t}$ forming the angle ϕ with respect to the real axis. The amplitude B and phase ϕ depend on the parameter κ. The angle $\phi = (\pi/2)(1 - \kappa^{1/2}/(\kappa+4)^{1/2})$. For $\kappa = 0$, $\phi = \pi/2$ and we recover the perpendicular stick; for $\kappa \to \infty$, $\phi \to 0$ and the angle of intersection decreases to zero. In the second case, $a_t = 2\sqrt{\kappa(1-t)}$, the behavior of the trace is more complex. For $0 < \kappa < 4$ the trace forms a finite spiral in the upper half plane; for $\kappa = 4$ the trace has a glancing intersection with the real axis. For $4 < \kappa < \infty$ the trace hits the real axis in accordance with the Hölder condition discussed above.

Stochastic Loewner Evolution

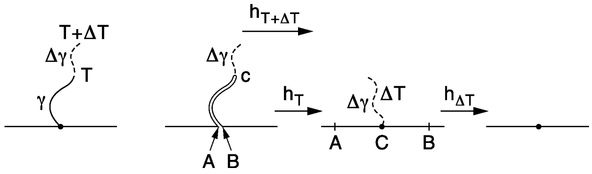
After these preliminaries we are in position to address stochastic Loewner evolution (SLE). The essential observation made by Oded Schramm [77] within the context of loop erased random walk was that the Markov and conformal properties of the measures or probability distributions for random curves generated by Loewner evolution imply that the random drive a_t must be proportional to an unbiased 1D Brownian motion.

Schramm's Theorem

The Loewner Equation (26) generates a non-crossing curve in the upper half plane $\mathbf{H}$ originating at the origin $\mathbf{O}$, given a continuous function a_t with initial value $a_0 = 0$. As a_t develops in time the tip of the curve z_t determined by the condition $g_t(z_t) = a_t$ traces out a curve or trace. In the case where a_t is a continuous random function the Loewner Equation (26) likewise becomes a stochastic equation of motion yielding a stochastic map $g_t(z)$. As a result the trace determined by $g_t(z_t) = a_t$ or $z_t = g_t^{-1}(a_t)$ is a random curve. The issue is to establish a contact between the exploration processes defining interfaces in the lattice models, the scaling limit of these curves, and the curves generated by SLE. In the scaling limit we thus invoke the two properties discussed above: i) the Markov property in Eq. (12) and ii) the conformal property in Eq. (16).

In order to demonstrate the surprising property that the Markov and conformal properties in combination imply that a_t must be a 1D Brownian motion we focus on chordal SLE which applies to a random curve or trace connecting two boundary points. Since the probability distribution or measure $P(\gamma)$ on the random curve γ using property ii) is assumed to be conformally invariant and since we by Riemann's theorem can map any simply connected domain to the upper half plane by mean of a conformal transformation, we are free to consider curves in the upper half plane from the origin $\mathbf{O}$ to ∞ parametrized with a time coordinate $0 < t < \infty$.

Imagine that we grow the curve from time $t = 0$ to time T driven by the function a_t, $0 < t < T$. The curve or trace is generated by the Loewner Equation (26) with boundary condition $a_0 = 0$ and the trace z_t by $g_t(z_t) = a_t$ or $z_t = g_t^{-1}(a_t)$. With the chosen time parametrization we have $g_t(z) \sim z + 2t/z$ for $z \to \infty$ in the upper half plane. The map g_t thus uniformizes the trace, i. e., the tip z_t is mapped to a_t on the real axis in the w-plane. In order to invoke the Markov property we let the curve grow the time increment ΔT corresponding to the curve segment $\Delta\gamma$. The Markov property then implies that the distribution on $\Delta\gamma$ conditioned on the distribution on γ is the same as the distribution on $\Delta\gamma$ in the cut domain $\mathbf{H} \setminus \gamma$, i. e., the domain with the curve γ deleted; this stage is illustrated in Fig. 7.



Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems, Figure 7
The figure depicts the construction in the derivation of SLE. The first step implements the Markov property by turning the curve γ into a cut. Subsequently, the conformal transformation h_T maps γ back to the origin. Finally, the map $h_{\Delta T}$ maps the segment $\Delta\gamma$ to the origin. The complete process is also implemented by $h_{T+\Delta T}$. The combination of the Markov property and conformal invariance implies that a_t performs a Brownian motion

Next, in order to implement the conformal property we shift the image by a_T in such a way that the curve segment $\Delta\gamma$ again starts at the origin $\mathbf{O}$. This is achieved by using the map $h_T = g_T - a_T$ which since $g_T(z_T) = a_T$ maps the tip z_T back to the origin; this construction is also depicted in Fig. 7. Moreover, the asymptotic behavior of h_T for large z is $h_T(z) \sim z - a_T - 2T/z$. Since the measure by assumption is unchanged under the conformal transformation h_T we infer that $\Delta\gamma$ growing the time ΔT from the origin has the same distribution as the segment $\Delta\gamma$ grown from time T to time $T + \Delta T$ conditioned on the segment γ grown up to time T. Moreover, since the segment γ from $\mathbf{O}(AB)$ to C subject to the Markov property has become part of the boundary, as shown in Fig. 7, we also infer that the measure on $\Delta\gamma$ is independent of the measure on γ. Finally, applying $h_{\Delta T}$ we map the segment $\Delta\gamma$ to the origin as a common reference point as indicated in Fig. 7.

Since the random curves are determined by the random maps g_t and h_t driven by the random function a_t the issue is how to transfer the properties of the measure on the curve determined by the Markov and conformal properties to the measure on the random driving function a_t.

In order to combine the Markov properties arising from the analysis of the lattice models and the conformal invariance pertaining to critical random curves, we carry out the following steps. First we grow the curve γ from the origin $\mathbf{O}$ to the tip z_T. Implying conformal invariance the curve is then uniformized back to the origin by means of h_T. The next step is to grow the curve segment $\Delta\gamma$ in time ΔT. This segment is subsequently absorbed by means of the map $h_{\Delta T}$. According to the Markov property the distribution of $\Delta\gamma$ from $\mathbf{O}$ to $z_{\Delta T}$ is the same as the distribution of $\Delta\gamma$ grown from T to $T + \Delta T$ conditioned on γ grown from 0 to T. Since the curve γ is determined by the map h_t the distribution is reflected in h_t. In particular, the stochastic properties of the curve is transferred to the random function a_t generating the curve by the Loewner evolution. The last step is now to observe that absorbing the segment $\Delta\gamma$ from **O** to $z_{\Delta T}$ by means of $h_{\Delta T}$ is the same transformation as first applying the inverse map h_T^{-1} followed by the map $h_{T+\Delta T}$; in both cases the end result is the absorption of the initial curve $\gamma + \Delta\gamma$, see Fig. 7. As regards the measure or distribution we have the equivalence $h_{\Delta T}(z) \approx h_{T+\Delta T}(h_T^{-1}(z))$. Using the asymptotic form $h_t(z) \sim z - a_t - 2t/z$ we obtain $a_{T+\Delta T} - a_T \approx a_{\Delta T}$; note that $\approx$ indicates identical distributions or measures.

In conclusion, the Markov property in combination with conformal invariance implies that $a_{T+\Delta T} - a_T$ is distributed like $a_{\Delta T}$ (stationarity) and that $a_{\Delta T}$ and $a_{\Delta T'}$ are independently distributed for non-overlapping time intervals ΔT and $\Delta T'$ (Markov property). Referring to Sect. "Scaling", A on Brownian motion as expressed in Eqs. (3) and (4), i. e., stationarity and independence, we infer that a_t is proportional to a Brownian motion of arbitrary strength κ, i. e., $a_t = \sqrt{\kappa} B_t$. Note that the reflection symmetry $x \to -x$ holding in the present context rules out a bias or drift in the 1D Brownian motion [13,27].

This is the basic conclusion reached by Schramm in the context of loop erased random walk. Driving Loewner evolution by means of 1D Brownian motion with different diffusion coefficient or strength κ we generate a one-parameter family of conformally invariant or scale invariant non-crossing random curves in the plane.

SLE Properties

Stochastic Loewner evolution is determined by the nonlinear stochastic equation of motion

$$\frac{\mathrm{d}g_t}{\mathrm{d}t} = \frac{2}{g_t - a_t} , \quad a_t = \sqrt{\kappa} B_t . \tag{27}$$

In the course of time a_t performs a 1D Brownian motion on the real axis starting at the origin $a_0 = 0$. a_t is a random continuous function of t and distributed according to $\sqrt{\kappa} B_t$. More precisely, a_t is given by the Gaussian distribution

$$P(a, t) = (2\pi t)^{-1/2} \exp[-a^2/2\kappa t] , \tag{28}$$

with correlations

$$\left\langle (a_t - a_s)^2 \right\rangle = \kappa |t - s| . \tag{29}$$

First we notice that a constant shift of the drive a_t, $a_t \to a_t + b$, is readily absorbed by a corresponding shift of the map, $g_t \to g_t + b$. Moreover, using the scaling property of Brownian motion, $B_{\lambda^2 t} = \lambda B_t$, following from e. g. Eq. (5), we have $a_{\lambda^2 t} = \lambda a_t$ and we conclude from Eq. (27) that $g_t(z)$ has the same distribution as $(1/\lambda) g_{\lambda^2 t}(\lambda z)$, i. e., $g_{\gamma^2 t}(\gamma z) \approx \gamma g_t(z)$. Note that this dilation invariance is consistent since the origin $z = 0$ and $z = \infty$, the endpoints of curves, are preserved. Note also that the strength of the drive κ is an essential parameter which cannot be scaled away.

Curves – Hulls – Bessel Process

For vanishing drive, $\kappa = 0$, the SLE yields a non random vertical line from $z = 0$, i. e., the growing stick discussed in Subsect. "Growing Stick". As we increase κ the curve becomes random with excursions to the right and to the left in the upper half plane. Up to a critical value of κ the random curve is simple; i. e., non-touching or non self-intersecting. At a critical value of κ the Brownian drive is so strong that the curve begins to intersect itself and the real axis. These intersection take place on all scales since the curve is self-similar or scale invariant. Denoting the curve by γ_t we observe that since Riemann's theorem uniformizing $\mathbf{H} \setminus \gamma_t$ to $\mathbf{H}$ only applies to a simply connected domain, the regions enclosed by the self-intersections do not become uniformized but are effectively removed from $\mathbf{H}$. The curve γ_t together with the enclosed parts is called the hull $\mathbf{K_t}$ and the mapping theorem applies to $\mathbf{H} \setminus \mathbf{K_t}$.

In order to analyze the critical value of κ we consider the stochastic equation for $h_t(z) = g_t(z) - a_t$. From Eq. (27) it follows that

$$\frac{\mathrm{d}h_t}{\mathrm{d}t} = \frac{2}{h_t} + \xi_t , \tag{30}$$

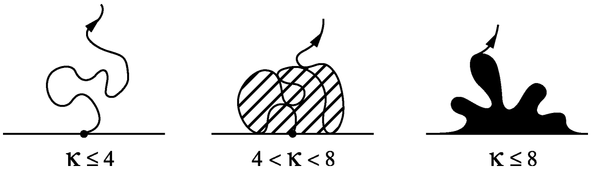
where $\xi_t = -\mathrm{d}a_t/\mathrm{d}t$ is white noise with correlations $\langle \xi_t \xi_s \rangle = \kappa\delta(t - s)$.

The nonlinear complex Langevin Equation (30) maps the tip of the curve z_t back to the origin. Likewise $\mathbf{H} \setminus \gamma_t$ is mapped onto $\mathbf{H}$. A point x on the real axis is mapped to $x_t = h_t(x)$ where x_t satisfies

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = \frac{2}{x_t} + \xi_t . \tag{31}$$

The Langevin Equation (31) is known as the Bessel equation and governs the radial distance R from the origin of a Brownian particle in d dimensions.

Introducing $R = (\sum_{i=1}^d B_i^2)^{1/2}$ where B_i, $i = 1, \dots d$, is a 1D Brownian motion with distribution $P(B, t) = (2\pi\kappa t)^{-1/2} \exp[-B^2/2\kappa t]$, we find $P(R, t) \propto (2\pi\kappa t)^{-d}$



Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems, Figure 8
The figures depict the phases of SLE. For $\kappa \leq 4$ the SLE trace is a simple non-intersecting scale invariant random curve from the origin to infinity with a fractal dimension between 1 and 3/2. For $4 < \kappa \leq 8$ the SLE curve is self-intersecting on all scales and also intersects the real axis on all scales. The curve together with the enclosed regions, the hull, eventually exhausts the upper half plane. The scale invariant hull has a fractal dimension ranging between 3/2 and 2. For $\kappa \geq 8$ the fractal dimension of the hull locks onto 2 and the scale invariant hull is dense and plane-filling

$R^{d-1}\exp[-R^2/2\kappa t]$ satisfying the Fokker–Planck equation $\partial P/\partial t = (\kappa/2)\partial^2 P/\partial R^2 - (\kappa(d-1)/2)\partial(P/R)/\partial R$, corresponding to the Langevin equation

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \kappa\frac{d-1}{2R} + \xi_t\,. \tag{32}$$

For $d \leq 2$ Brownian motion is recurrent, i. e., the particle returns to the origin $R = 0$, for $d > 2$ the particle goes off to infinity. Referring to Eq. (31) and setting $\kappa(d-1)/2 = 2$ we obtain $\kappa = 4/(d-1)$, i. e., $R \to \infty$ for $\kappa < 4$ and $R \to 0$ for $\kappa > 4$.

Since the tip of the curve z_t is mapped to a_t, i. e., $h(z_t) = 0$, the case $x_t \to \infty$ for $\kappa < 4$ corresponds to a curve never intersecting the real axis, i. e., the curve is simple. For $\kappa > 4$ we have $x_t \to 0$ corresponding to the case where the tip z_t intersects the real axis forming a hull. Since the curve is self-similar the intersections takes place on all scales and eventually the whole upper half plane is engulfed by the hull.

The marginal value $\kappa = 4$ can also be inferred from a simple heuristic argument [27]. For small κ we can ignore the noise in Eq. (31)and the particle is repelled according to the solution $x_t^2 \sim 4t$. For large noise we ignore the nonlinear term and the noise can drive x_t to zero; we have $x_t^2 \sim \kappa t$. The balance is obtained for $\kappa = 4$.

In conclusion, for $\kappa \leq 4$ the curve is simple, for $\kappa > 4$ the curve intersects itself and the real axis infinitely many times on all scales, eventually the hull swallows the whole plane. For large κ the trace turns out to be plane-filling. The two cases $\kappa \leq 4$ and $\kappa > 4$ are depicted in Fig. 8.

Fractal Dimension

The SLE random curves are fractal. An important issue is thus the determination of the fractal dimension D in terms of the Brownian strength or SLE parameter κ. It has been shown that [15,16,74]

$$D = 1 + \frac{\kappa}{8} \quad \text{for } 0 \leq \kappa \leq 8\,, \tag{33}$$

for $\kappa \geq 8$ the fractal dimension locks onto 2 and the SLE curve is plane-filling.

In order to illustrate a typical SLE calculation we follow Cardy [27] in a heuristic derivation of Eq. (33). In order to evaluate D and in accordance with its definition [32,68] the standard procedure is to cover the object with disks of size ε and follow how the number of disks $N(\epsilon)$ of size ε scales with ε for small ε, i. e., $N(\epsilon) \propto \epsilon^{-D}$. However, since the SLE curve is random the argument has to be rephrased. Alternatively, we consider a disc of size ε located at a fixed position z and ask for the probability $P(z,\epsilon)$ that the SLE curve crosses the disk. The number of disks covering an area A is $N_A = A/\epsilon^2$, i. e., $P \propto \epsilon^{-D}/N_A$, and we infer $P(z,\epsilon) \propto \epsilon^{2-D}$, where D is the fractal dimension. Incorporating the Markov property we subject the curve to an infinitesimal conformal transformation, $h_{\delta t} = g_{\delta t} - a_{\delta t}$, transforming the point z to $w = g_{\delta t}(z) - a_{\delta t}$; moreover, all lengths are scaled by $|h'_{\delta t}(z)|$ [1]. Setting $z = x + iy$ and $w = x' + iy'$ we obtain expanding Eq. (30) $x' + iy' = 2\delta t/(x+iy) - \sqrt{\kappa}\delta B_t$ or $x' = x + 2x\delta t/(x^2+y^2) - \sqrt{\kappa}\delta B_t$, $y' = y - 2y\delta t/(x^2+y^2)$, $\epsilon' = (1 - |h_{\delta t}(z)|)\epsilon$, and $|h_{\delta t}| = 2(x^2-y^2)/(x^2+y^2)^2$. By conformal invariance the probability measure $P(x,y,\epsilon)$ is unchanged and we infer

$$P(x,y,\epsilon) = \langle P(x',y',\epsilon')\rangle_{\delta B}\,, \tag{34}$$

where we have averaged over the Brownian motion referring to the initial part of the curve which has been eliminated by the map $h_{\delta t}$. Expanding Eq. (34) to first order in δt and noting that $\langle(\delta B_t)^2\rangle = \delta t$ we arrive at a partial differential equation for $P(x,y,\epsilon)$

$$\left(\frac{2x}{x^2+y^2}\frac{\partial}{\partial x} - \frac{2y}{x^2+y^2}\frac{\partial}{\partial y} + \frac{\kappa}{2}\frac{\partial^2}{\partial x^2} - \frac{2(x^2-y^2)}{(x^2+y^2)^2}\epsilon\frac{\partial}{\partial\epsilon}\right)P = 0\,. \tag{35}$$

Since $P(x,y,\epsilon) \propto \epsilon^{2-D}$ we have $\epsilon\partial P/\partial\epsilon = (2-D)P$ and the determination of D is reduced to an eigenvalue problem. By inspection one finds

$$P \propto \epsilon^{1-\kappa/8} y^{(\kappa-8)^2/8\kappa}(x^2+y^2)^{(\kappa-8)/2\kappa}\,, \tag{36}$$

and we identify the fractal dimension $D = 1 + \kappa/8$ for $\kappa < 8$; for $\kappa > 8$ another solution yields $D = 2$.

Results and Discussion

Stochastic Loewner evolution based on Eq. (27) generates conformally invariant non-crossing random curves in the upper half plane starting at the origin and going off to infinity. This is the case of chordal SLE, where the random curve connects two boundary points (the origin **O** and infinity ∞). Another case is radial SLE for random curves connecting a boundary point and an interior point in a simply connected domain [13,27,43]. Radial SLE is governed by another stochastic equation and will not be discussed here.

Phase Transitions – Locality – Restriction – Duality

Phase Transitions SLE exhibits two phase transitions; for $\kappa = 4$ and $\kappa = 8$. For $0 < \kappa \leq 4$ the random curve is non-intersecting, i. e., a simple random continuous curve from **O** to ∞. For $4 < \kappa \leq 8$ the curve is self-intersecting on all scales. The curve together with the excluded regions form a hull which in the course of time absorbs the upper half plane. For κ just above 4 the half plane is eventually absorbed but the trace does not visit all regions, i. e., the hull is not dense. As we approach $\kappa = 8$ the trace becomes more dense and the hull becomes plane-filling. This is also reflected in the fractal dimension $D = 1 + \kappa/8$. For $\kappa > 8$ the hull is plane-filling, i. e., $D = 2$. As we increase the strength of the Brownian drive further the excursions of the trace to the right and left in the upper half plane become more pronounced and the hull becomes vertically compressed. These results have been obtained by Rohde and Schramm [74] and Lawler et al. [57]. The various phases of SLE are depicted in Fig. 8. In Fig. 9 we have depicted numerical renderings of SLE traces for various values of κ (with permission from V. Beffara, http://www.umpa.ens-lyon.fr/~vbeffara/simu.php).

Locality – Restriction In addition to the phase transitions at $\kappa = 4$ and $\kappa = 8$, there are special values of κ where SLE shows a behavior characteristic of the scaling limit of specific lattice models: The locality property for $\kappa = 6$ and the restriction property for $\kappa = 8/3$. The issue here is the influence of the boundary on the SLE trace.

To illustrate the locality property, consider for example the SLE trace originating at the origin and purporting to describe the scaling limit of a domain wall in the lattice model. Due to the long range correlations at the critical point it is intuitively clear that a deformation of the boundary, e. g., a bulge **L** on the real axis to the right of the origin, will influence the trace and push it to the left. A detailed analysis show that only for $\kappa = 6$ is the trace independent of a change of the boundary, i. e., the trace does not feel the boundary until it encounters a boundary point [59,63]. Returning to the lattice models the locality property for $\kappa = 6$ applies specifically to the percolation case where the interface generated by the exploration process is governed by a local rule and the model has a geometric phase transition.

The restriction property is less obvious to visualize but basically states that the distribution of traces conditioned not to hit a bulge **L** on the real axis away from the origin is the same as the distribution of traces in the domain where **L** is part of the boundary, i. e., in the domain $\mathbf{H} \setminus \mathbf{L}$. Analysis shows that the restriction property only applies in the case for $\kappa = 8/3$. Among the lattice models only the scaling limit of self-avoiding random walk (SAW), where the measure is uniform, conforms to the restriction property and thus corresponds to $\kappa = 8/3$ [63].

Duality For $\kappa > 4$ the SLE generates a hull of fractal dimension $D > 3/2$. The boundary, external perimeter, or frontier of the hull is again a simple conformally invariant random curve characterized by the fractal dimension $\bar{D}$. Using methods from 2D quantum gravity Duplantier [31] has proposed the relationship,

$$(D-1)(\bar{D}-1) = \tfrac{1}{4}\,, \tag{37}$$

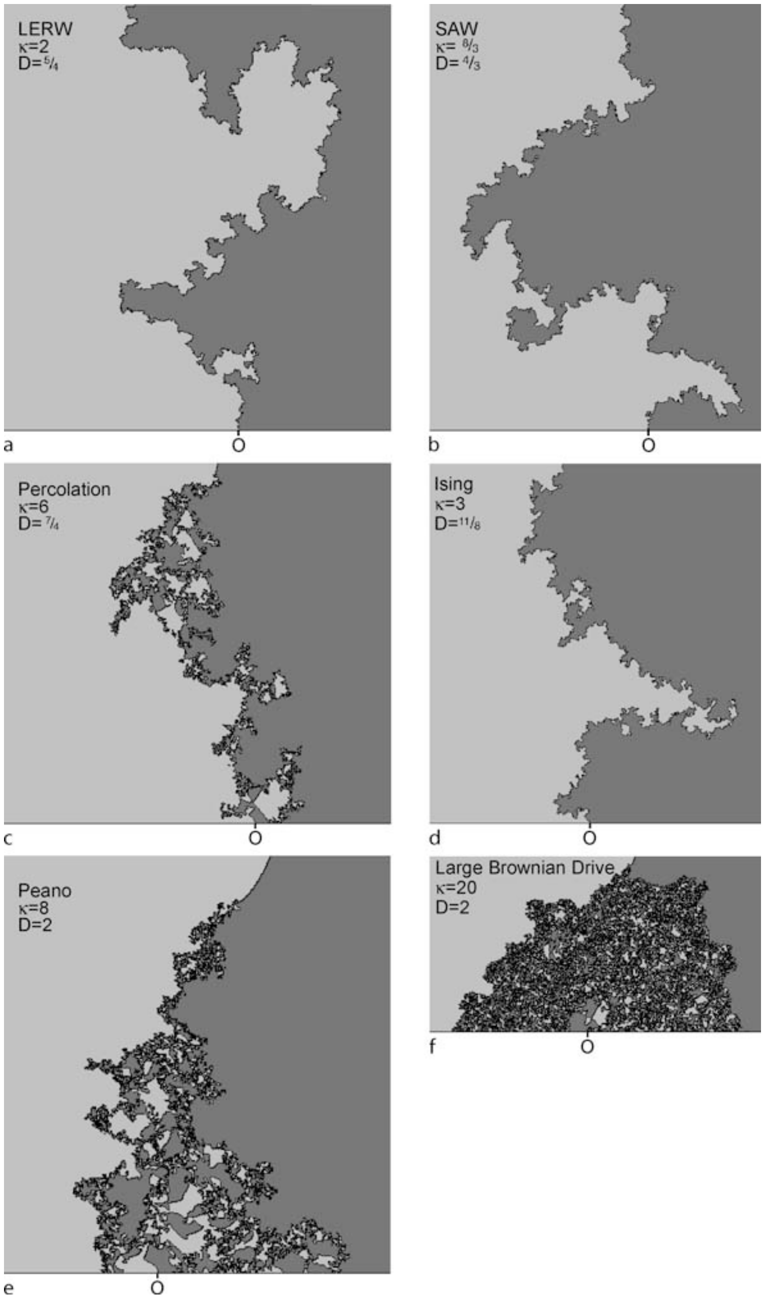
between the fractal dimension of the hull and its frontier. This result has been proved by Beffara for $\kappa = 6$, i. e., the percolation case [16]. Inserting in Eq. (33) we obtain for the corresponding SLE parameter the duality relation

$$\kappa\bar{\kappa} = 16\,. \tag{38}$$

Loop Erased Random Walk (LERW)

Whereas the scaling limit of random walk, i. e., Brownian motion, does not fall in the SLE category because of self-crossings rendering Riemann's mapping theorem inapplicable, variations of Brownian motion are described by SLE.

Loop erased random walk (LERW) where loops are removed along the way is by construction self-avoiding and was introduced as a simple model of a self-avoiding random walk. LERW was studied by Schramm in his pioneering work [77]. LERW has the Markov property and has been proved to be conformally invariant in the scaling limit and described by SLE for $\kappa = 2$ [57]. According to Eq. (33) LERW has the fractal dimension $D = 5/4$ [67]. Also, since $\kappa < 4$ LERW is non-intersecting. A simulation of LERW based on SLE is shown in Fig. 9a.



Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems, Figure 9
We depict a numerical rendering of SLE for a variety of κ values. In a we show loop erased random walk (LERW) for $\kappa = 2$ with fractal dimension $D = 5/4$. In b we illustrate the case of self-avoiding random walk (SAW) for $\kappa = 8/3$ and fractal dimension $D = 4/3$; both LERW and SAW have $\kappa > 4$ and are simple scale invariant random curves. In c we depict site percolation for $\kappa = 6$ with fractal dimension $D = 7/4$. Since $\kappa > 4$ the percolation case is self-intersecting and duality implies that the boundary or frontier of the hull is described by a SLE curve for $\kappa = 16/6 = 8/3$, i. e., the case of SAW. In d we show the Ising case for $\kappa = 3$ and fractal dimension $D = 11/8$. In e we depict the limiting case $\kappa = 8$ and fractal dimension $D = 2$. The hull is dense and plane-filling. The frontier of the hull corresponds to the SLE case $\kappa = 16/8 = 2$, i. e., the case of LERW. The so-called uniform spanning tree (UST) has the same properties as LERW and the SLE case for $\kappa = 8$ can thus be thought of as a random plane filling Peano curve wrapping around the UST. Finally, in f we show the SLE trace and hull for $\kappa = 20$ and $D = 2$. Because of the large Brownian excursions the plane-filling hull is vertically compressed (with permission from V. Beffara: http://www.umpa.ens-lyon.fr/~vbeffara/simu.php)

Self-Avoiding Random Walk (SAW)

Self-avoiding random walk (SAW) is a random walk conditioned not to cross itself. SAW has been used to model polymers in a dilute solution and has a uniform probability measure. Since SAW satisfies the restriction property it is conjectured in the scaling limit to fall in the SLE class with $\kappa = 8/3$ [46,47,48,61], yielding the fractal dimension $D = 4/3$. We note that Flory's mean field theory [30] for the size R of a polymer composed of N links (monomers) scales like $R \sim N^{\nu}$, where $\nu = 3/(2+d)$ for $d \leq 4$. By a box covering we infer $N \sim R^{D}$ where D is the fractal dimension, i. e., $D = (2+d)/3$. In $d = 2$ we obtain $D = 4/3$ in accordance with the SLE result. SLE induced SAW in the scaling limit is shown in Fig. 9b.

Percolation

The scaling limit of site percolation was conjectured by Schramm [77] to fall in the SLE class for $\kappa = 6$. Subsequently, the scaling limit of site percolation on a triangular lattice has been proved by Smirnov [78,80]. Percolation exhibits a geometrical phase transition. In the exploration process defining a critical interface the rule for propagation is entirely local. The lack of stiffness as for example in the Ising case to be discussed below results in a strongly meandering path winding back and in the scaling limit intersecting earlier part of the path. Since $\kappa > 4$ the path together with the enclosed part, i. e., the hull, eliminates the whole plane in the course of time. As discussed above the locality property is specific to percolation and yields $\kappa = 6$. The fractal dimension of the percolation interface is according to Eq. (33) $D = 7/4$. We note that D is close to 2, i. e., the percolation interface nearly covers the plane densely. A series of new results and a proof of Cardy's conjectured formula for the crossing probability have appeared; we refer to [13,27,43] for details. Using the duality relation (38) the frontier of the percolation hull is a simple SLE curve for $\kappa = 8/3$, corresponding to SAW. In Fig. 9c we have depicted a SLE generated percolation interface.

Ising Model – $O(n)$ Models

The Ising model in Eq. (6) is a special case of the $O(n)$ model defined by the Hamiltonian

$$H = -J \sum_{\langle ij \rangle} \vec{\sigma}_i \vec{\sigma}_j \,, \tag{39}$$

where $\vec{\sigma}_i = (\sigma_1, \ldots \sigma_n)$ is an n-component unit vector associated with the site i. For $n = 1$ we recover the Ising model, $n = 2$ is the XY-model [29], and $n = 3$ the Heisenberg model.

By means of the Fortuin–Kasteleyn (FK) transformation based on a high temperature expansion the configurations of the $O(n)$ model can be described by clusters or graphs on a dual lattice [35]. The crossing domain wall in Fig. 1 is thus a special case of a FK graph if we interpret the representation as a triangular Ising model. It has been conjectured that n is related to the SLE parameter κ by

$$n = -2\cos(4\pi/\kappa) \quad \text{for } 8/3 \leq \kappa \leq 4\,. \tag{40}$$

In the Ising case $n = 1$ and we have $\kappa = 3$ yielding the fractal dimension $D = 11/8$ for the Ising domain wall [84]. Since $\kappa < 4$ the Ising domain wall is non-intersecting. Unlike the percolation case, the Ising interface is stiffer due to the interaction. We also note that the scaling limit of spin cluster boundaries in the Ising model recently has been proved to correspond to SLE for $\kappa = 3$ [79]. The interface is shown in Fig. 9d.

SLE – Conformal Field Theory

Whereas conformal field theory (CFT) is based on the concept of a local field $\phi(r)$ and its correlations and therefore only accesses the underlying geometry indirectly through field correlations, SLE directly produces conformally invariant geometrical objects. A major issue is therefore the connection between CFT and SLE [6,7,26]. In CFT the central charge c plays an important role in delimiting the universality classes of the variety of lattice models yielding conformal field theories in the scaling limit. Percolation thus corresponds to the central charge $c = 0$, whereas the Ising model is associated with the central charge $c = 1/2$. It has been conjectured that the connection between the SLE parameter κ and the central charge c is given by

$$c = \frac{(6-\kappa)(3\kappa-8)}{2\kappa} = 1 - 6\frac{(\kappa-4)^2}{4\kappa}\,. \tag{41}$$

We note that $c < 1$ and, moreover, invariant under the duality transformation $\kappa \to 16/\kappa$.

SLE – 2D Turbulence

There is an interesting application of SLE ideas in the context of 2D turbulence. The issue here is to analyze conformal invariance by comparing the statistical properties of geometrical shapes like domain walls with SLE traces with the view of determining the SLE parameter κ and the corresponding universality class.

In 3D turbulence is governed by the incompressible Navies–Stokes equation for the velocity field. Since the viscosity is only effective at small length scales 3D turbulence is characterized by a cascade of kinetic energy $(1/2)v^2$ from large scales (driving scale) to small scales (dissipation scale). In the inertial regime the energy spectrum $E(k)$ (k is the wavenumber) is characterized by the celebrated Kolmogorov 5/3 law [46], $E(k) \propto k^{-5/3}$, indicating an underlying scale invariance in turbulence.

In 2D the cascade picture is different. Since both kinetic energy and squared vorticity (enstrophy) are conserved in the absence of dissipation and forcing, two cascades coexist [52,53]. A direct cascade to small scales for the squared vorticity $\omega^2 = (\nabla \times v)^2$ with scaling exponent -3 and an inverse cascade to larger scales for the kinetic energy $(1/2)v^2$ with Kolmogoroff scaling exponent $-5/3$. The system is thus characterized by a fine scale vorticity structure together with a large scale velocity structure. Moreover, we can assume that the vorticity structure is equipartitioned, i.e, in equilibrium.

In order to investigate whether the scale invariance of the small scale vorticity structure can be extended to conformal invariance Bernard et al. [17] have considered the statistics of the boundaries of vorticity clusters. By comparing the zero-vorticity isolines with SLE traces they find that cluster boundaries fall in the universality class corresponding to $\kappa = 6$, i. e., the case of percolation. Since 2D turbulence is a driven nonequilibrium system, this observation is very intriguing in particular since the correlations between vortices are long-ranged. A similar analysis [19] of the isolines in the inverse cascade in surface quasi-geostrophic turbulence corresponds to $\kappa = 4$, i. e., from Eq. (40) the domain walls in the equilibrium XY model for $n = 2$. For comments on the application of SLE in turbulence we refer to Cardy [28].

SLE – 2D Spin Glass

It is a standing issue whether conformal field theory can be applied to disordered systems, in particular systems with quenched disorder. In recent work Amoruso et al. [3] and Bernard et al. [18] have considered zero temperature domain walls in the Ising spin glass [34]; see also [33]. The Ising spin glass is an equilibrium system with quenched disorder. The system is described by the Hamiltonian $H = -\sum_{\langle ij\rangle} J_{ij}\sigma_i\sigma_j$, where the random exchange constants J_{ij} are picked from a Gaussian distribution with zero mean. The glass transition is at $T = 0$ and the system has a two-fold degenerate ground state. Inducing a scale invariant domain wall between the two ground states and comparing with an SLE trace, it is found that both the Markov and conformal properties are obeyed and that the universality class corresponds to $\kappa \approx 2.3$.

Further Remarks

In this discussion we have left out several topics which have played an important role in the development and applications of SLE. We mention some of them below.

There is an interesting connection between LERW and the so-called uniform spanning tree (UST) [13,43,57,77]. A spanning tree is a collection of vertices and edges which form a tree, i. e., without loops or cycles. A uniform spanning tree is a random spanning tree picked among all possible spanning trees with equal probability. Consider the unique path between two vertices on a UST. Since the path lives on a tree it is by construction non-crossing and it turns out that it has the same distribution as LERW. The winding random curve enclosing the UST can be visualized as a random plane-filling Peano curve. In the scaling limit the Peano curve is described by SLE for $\kappa = 8$ with fractal dimension $D = 2$.

The q-state Potts model [87] constitutes a generalization of the Ising model; here the lattice variable takes q values. The model is defined by the Hamiltonian $H = -J\sum_{\langle ij\rangle} \delta_{\sigma_i\sigma_j}$, where $\sigma_i = 1, \ldots q$; the Ising model obtains for $q = 2$. Applying the high temperature FK representation the configurations can be represented by loops and domain walls. From considerations involving the fractal dimension [76] it has been conjectured that domain walls in the scaling limit of the Potts model fall in the SLE category for $q = 2 + 2\cos(8\pi/\kappa)$, where $4 \leq \kappa \leq 8$. For $q = 2$ we recover the Ising case for $\kappa = 3$. In the limit $q \to 0$, the graph representation is equivalent to the uniform spanning tree described by SLE for $\kappa = 8$. For a numerical study of the three-state Potts model and its relation to SLE consult [36].

Standard SLE is driven by 1D Brownian motion producing a fractal curve. Ruskin et al. [75] have considered the case of adding a stable Lévy process with shape parameter α to the Brownian motion. Backing their analysis with numerics they find that the SLE trace branches and exhibit a 'phase transitions' related to self-intersections.

2D Brownian motion, the scaling limit of 2D random walk, is an incredibly complex fractal coil owing to the self-crossings on all scales. Although 2D Brownian motion because of self-crossing itself falls outside the SLE scheme, the outer frontier or perimeter of 2D random walk is a non-crossing and non-intersecting fractal curve which can be accessed by SLE. Verifying an earlier conjecture by Mandelbrot [68] it has been proven using SLE techniques [58] that the fractal dimension of the Brownian perimeter is $D = 4/3$, i.e, the same as the fractal dimension of self-avoiding random walk and the external perimeter of the percolation hull. Other characteristics of Brownian motion such as intersection exponents have also been obtained [59,60,62]; see also [66].

In recent work Zoia et al. [88] have considered the distribution of first passage times and distances along critical curves generated by SLE for different values of κ.

Future Directions

Stochastic Loewner evolution represents a major step in our understanding of fractal shapes in the 2D continuum limit. By combining the Markov property (stationarity) with conformal invariance SLE provides a minimal scheme for the generation of a one-parameter family of fractal curves. The SLE scheme also provides calculational tools which have led to a host of new results. SLE is a developing field and we can on the mathematical front anticipate progress and proofs of some yet unproved scaling limits, e. g., the scaling limit of the FK representation of the Potts model and the scaling limit of SAW.

On the more physical front many issues also remain open. First there is the fundamental issue of the connection between the hugely successful but non-rigorous CFT and SLE. Here progress is already under way. In a series of papers Bauer and Bernard [6,7,8,9,10,12] have shown how SLE results can be derived using CFT methods. Cardy [26] have considered a multiple SLE process and the connection to Dyson's Brownian process and random matrix theory. The analysis of the CFT–SLE connection still remains to be investigated further.

An obvious limitation of SLE is that it only addresses critical domain walls and not the full configuration of clusters and loops in for example the FK representation of the Potts model. In the case of critical percolation this problem has been addressed by Camia and Newman [21]. Another issue is how to provide SLE insight into spin correlations in the Potts or $O(n)$ models.

In the original formulation of SLE the Markov and conformal properties essentially requires a Brownian drive. It is clearly of interest to investigate the properties of random curves generated by other random drives. Such a program has been initiated by Ruskin et al. [75] who considered adding a Lévy drive to the Brownian drive; see also work by Kennedy [49,50].

Since the SLE trace lives in the infinite upper half plane the whole issue of finite size effects remain open. In ordinary critical phenomena the concept of a Kadanoff block construction and the diverging correlation length near the transition lead to a theory of finite size scaling and corrections to scaling which can be accessed numerically. It is an open problem how to develop a similar scheme for SLE.

In statistical physics it is customary and natural to associate a free energy to a domain wall and an interaction energy associated with several domain walls. These free energy considerations are entirely absent in the SLE framework which is based on conformal transformations. A major issue is thus: Where is the free energy in all this and how do we reintroduce and make use of ordinary physical considerations and estimates [69].

Acknowledgments

The present work has been supported by the Danish Natural Science Research Council.

Bibliography

1. Ahlfors LV (1966) Complex analysis: An introduction to the theory of analytical functions of one complex variable. McGraw-Hill, New York
2. Ahlfors LV (1973) Conformal invariance: Topics in geometric function theory. McGraw-Hill, New York
3. Amoruso C, Hartmann AK, Hastings MB, Moore MA (2006) Conformal invariance and stochastic Loewner evolution processes in two-dimensional Ising spin glasses. Phys Rev Lett 97(4):267202. arXiv:cond-mat/0601711
4. Ash RB, Doléans CA (2000) Probability and measure theory. Academic, San Diego
5. Bak P (1999) How nature works: The science of self-organized criticality. Springer, New York
6. Bauer M, Bernard D (2002) SLE$_\kappa$ growth processes and conformal field theory. Phys Lett B 543:135–138. arXiv:math.PR/0206028
7. Bauer M, Bernard D (2003) Conformal field theories of stochastic Loewner evolutions. Comm Math Phys 239:493–521. arXiv:hep-th/0210015
8. Bauer M, Bernard D (2003) SLE martingales and the Viasoro algebra. Phys Lett B 557:309–316. arXiv:hep-th/0301064
9. Bauer M, Bernard D (2004) CFTs of SLEs: The radial case. Phys Lett B 583:324–330. arXiv:math-ph/0310032
10. Bauer M, Bernard D (2004) Conformal transformations and the SLE partition function martingale. Ann Henri Poincare 5:289–326. arXiv:math-ph/0305061
11. Bauer M, Bernard D (2004) Loewner chains. arXiv:cond-mat/0412372
12. Bauer M, Bernard D (2004) SLE, CFT and zig-zag probabilities. Proceedings of the conference 'Conformal Invariance and Random Spatial Processes', Edinburgh, July 2003. arXiv:math-ph/0401019
13. Bauer M, Bernard D (2006) 2D growth processes: SLE and Loewner chains. Phys Rep 432:115–221
14. Baxter RJ (1982) Exactly solved models in statistical mechanics. Academic, London
15. Beffara V (2002) The dimension of SLE curves. arXiv:math.PR/0211322

16. Beffara V (2003) Hausdorff dimensions for SLE_6. Ann Probab 32:2606–2629. arXiv:math.PR/0204208
17. Bernard D, Boffetta G, Celani A, Falkovich G (2006) Conformal invariance in two-dimensional turbulence. Nat Phys 2:124–128
18. Bernard D, Le Doussal P, Middleton AA (2006) Are domain walls in 2D spin glasses described by stochastic Loewner evolutions. arXiv:cond-mat/0611433
19. Bernard D, Boffetta G, Celani A, Falkovich G (2007) Inverse turbulent cascades and conformally invariant curves. Phys Rev Lett 98:024501(4). arXiv:nlin.CD/0602017
20. Binney JJ, Dowrick NJ, Fisher AJ, Newman MEJ (1992) The theory of critical phenomena. Clarendon, Oxford
21. Camia F, Newman CM (2003) Continuum nonsimple loops and 2D critical percolation. arXiv:math.PR/0308122
22. Cardy J (1987) Conformal invariance. In: Domb C, Lebowitz JL (eds) Phase transitions and critical phenomena, vol 11. Academic, London
23. Cardy J (1993) Conformal field theory comes of age. Physics World, June, 6:29–33
24. Cardy J (1996) Scaling an renormalization in statistical physics. Cambridge University Press, Cambridge
25. Cardy J (2002) Conformal invariance in percolation, self-avoiding walks and related problems. Plenary talk given at TH-2002, Paris. arXiv:cond-mat/0209638
26. Cardy J (2003) Stochastic Loewner evolution and Dyson's circular ensembles. J Phys A 36:L379–L408. arXiv:math-ph/0301039
27. Cardy J (2005) SLE for theoretical physicists. Ann Phys 318:81–118. arXiv:cond-mat/0503313
28. Cardy J (2006) The power of two dimensions. Nat Phys 2:67–68
29. Chaikin PM, Lubensky TC (1995) Principles of condensed matter physics. Cambridge University Press, Cambridge
30. de Gennes PG (1985) Scaling concepts in polymer physics. Cornell University Press, Ithaca
31. Duplantier B (2000) Conformally invariant fractals and potential theory. Phys Rev Lett 84:1363–1367. arXiv:cond-mat/9908314
32. Feder J (1988) Fractals (physics of solids and liquids). Springer, New York
33. Fisch R (2007) Comment on conformal invariance and stochastic Loewner evolution processes in two-dimensional Ising spin glasses. arXiv:0705.0046
34. Fischer KH, Hertz JA (1991) Spin glasses. Cambridge University Press, Cambridge
35. Fortuin CM, Kasteleyn PW (1972) On the random cluster model. Physica 57:536–564
36. Gamsa A, Cardy J (2007) SLE in the three-state Potts model – a numerical study. arXiv:0705.1510
37. Gardiner CW (1997) Handbook of stochastic methods. Springer, New York
38. Gong S (1999) The Bieberbach conjecture. RI American 19. Mathematical Society. International, Providence
39. Gruzberg IA, Kadanoff LP (2004) The Loewner equation: Maps and shapes. J Stat Phys 114:1183–1198. arXiv:cond-mat/0309292
40. Jensen HJ (2000) Self-organized criticality: Emergent complex behavior in physical and biological systems. Cambridge University Press, Cambridge
41. Kadanoff LP (1966) Scaling laws for Ising models near T_c. Physics 2:263–271
42. Kadanoff LP, Berkenbusch MK (2004) Trace for the Loewner equation with singular forcing. Nonlinearity 17:R41–R54. arXiv:cond-mat/0402142
43. Kager W, Nienhuis B (2004) A guide to stochastic Loewner evolution and its application. J Stat Phys 115:1149–1229
44. Kager W, Nienhuis B, Kadanoff LP (2004) Exact solutions for loewner evolutions. J Stat Phys 115:805–822
45. Kauffman SA (1996) At home in the universe: The search for the laws of self-organization and complexity. Oxford University Press, Oxford
46. Kennedy T (2002) Monte Carlo tests of stochastic Loewner evolution predictions for the 2D self-avoiding walk. Phys Rev Lett 88(4):130601. arXiv:math.PR/0112246
47. Kennedy T (2004) Conformal invariance and stochastic Loewner evolution predictions for the 2D self-avoiding walk – Monte Carlo tests. J Stat Phys 114:51–78. arXiv:math.PR/0207231
48. Kennedy T (2005) Monte Carlo comparisons of the self-avoiding walk and SLE as parameterized curves. arXiv:math.PR/0510604v1
49. Kennedy T (2006) The length of an SLE – Monte Carlo studies. arXiv:math.PR/0612609v1
50. Kennedy T (2007) Computing the Loewner driving process of random curves in the half plane. arXiv:math.PR/0702071v1
51. Kolmogorov AN (1941) Dissipation of energy in the locally isotropic turbulence. Dokl Akad Nauk SSSR 30:9–13. (Reprinted in Proc Royal Soc Lond A 434:9–13 (1991))
52. Kraichnan RH (1967) Inertial ranges in two-dimensional turbulence. Phys Fluids 10:1417–1423
53. Kraichnan RH, Montgomery D (1980) Two-dimensional turbulence. Rep Prog Phys 43:567–619
54. Landau LD, Lifshitz EM (1959) Theory of elasticity. Pergamon, Oxford
55. Lawler GF (2004) ICTP Lecture Notes Series. 17:307–348
56. Lawler GF (2005) Conformally invariant processes in the plane. In: Mathematical Surveys and Monographs, vol 114. AMS, Providence
57. Lawler GF, Schramm O, Werner W (2001) Conformal invariance of planar loop-erased random walks and uniform spanning trees. Ann Prob 32:939–995. arXiv:math.PR/0112234
58. Lawler GF, Schramm O, Werner W (2001) The dimension of the planar Brownian frontier is 4/3. Math Res Lett 8:401–411. arXiv:math.PR/00010165
59. Lawler GF, Schramm O, Werner W (2001) Values of Brownian intersections exponents I: Half plane exponents. Acta Math 187:237–273. arXiv:math.PR/9911084
60. Lawler GF, Schramm O, Werner W (2001) Values of Brownian intersections exponents II: Plane exponents. Acta Math 187:275–308. arXiv:math.PR/0003156
61. Lawler GF, Schramm O, Werner W (2002) On the scaling limit of planar self-avoiding walk. Fractal geometry and application, a jubilee of Benoit Mandelbrot, Part 2, 339–364, Proc. Sympos. Pure Math., 72, Part 2, Amer. Math. Soc., Providence, RI, 2004. arXiv:math.PR/0204277
62. Lawler GF, Schramm O, Werner W (2002) Values of Brownian intersections exponents III: Two-sided exponents. Ann Inst Henri Poincare 38:109–123. arXiv:math.PR/0005294
63. Lawler GF, Schramm O, Werner W (2003) Conformal restriction: The chordal case. J Amer Math Soc 16:917–955. arXiv:math.PS/0209343

64. Löwner K (Loewner C) (1923) Untersuchungen über schlichte konforme Abbildungen des Einheitskreises. I Math Ann 89:103–121
65. Ma S-K (1976) Modern theory of critical phenomena. Frontiers in Physics, vol 46. Benjamin, Reading
66. Mackenzie D (2000) Taking the measure of the wildest dance on earth. Science 290:1883–1884
67. Majumbar SN (1992) Exact fractal dimension of the loop-erased self-avoiding walk in two dimensions. Phys Rev Lett 68:2329–2331
68. Mandelbrot B (1987) The fractal geometry of nature. Freeman, New York
69. Moore M (2007) private communication
70. Nicolis G (1989) Exploring complexity: An introduction. Freeman, New York
71. Nienhuis B (1987) Coulomb gas formulation of two-dimensional phase transitions. In: Domb C, Lebowitz JL (eds) Phase transitions and critical phenomena, vol 11. Academic, London
72. Pfeuty P, Toulouse G (1977) Introduction to the renormalization group and to critical phenomena. Wiley, New York
73. Reichl LE (1998) A modern course in statistical physics. Wiley, New York
74. Rohde S, Schramm O (2001) Basic properties of SLE. Ann Math 161:879–920. arXiv:mathPR/0106036
75. Rushkin I, Oikonomou P, Kadanoff LP, Gruzberg IA (2006) Stochastic Loewner evolution driven by Levy processes. J Stat Mech (2006) 01:P01001. arXiv:cond-mat/0509187
76. Saleur H, Duplantier B (1987) Exact determination of the percolation hull exponent in two dimensions. Phys Rev Lett 58:2325–2328
77. Schramm O (2000) Scaling limit of loop-erased random walks and uniform spanning trees. Israel J Math 118:221–288. arXiv:math.PR/9904022
78. Smirnov S (2001) Critical percolation in the plane: Conformal invariance, Cardy's formula, scaling limits. C R Acad Sci Paris Ser I Math 333(3):239–244
79. Smirnov S (2006) Towards conformal invariance of 2D lattice models. Proceedings of the international congress of mathematicians (Madrid, August 22–30, 2006). Eur Math Soc 2:1421–1451
80. Smirnov S, Werner W (2001) Critical exponents for two-dimensional percolation. Math Res Lett 8:729–744
81. Stanley HE (1987) Introduction to phase transitions and critical phenomena. Oxford University Press, Oxford
82. Stauffer D, Aharony A (1994) Introduction to percolation theory. CRC, Boca Raton
83. Strogatz S (2003) Sync: The emerging science of spontaneous order. Hyperion, New York
84. Vanderzande C, Stella AL (1989) Bulk, surface and hull fractal dimension of critical Ising clusters in $d = 2$. J Phys A: Math Gen 22:L445–L451
85. Werner W (2004) Random planar curves and Schramm–Loewner evolutions. Springer Lecture Notes in Mathematics 1840:107–195. arXiv: math.PR/0303354
86. Wilson KG, Kogut J (1974) The renormalization group and the ε expansion. Phys Rep 12:75–199
87. Wu FY (1982) The Potts model. Rev Mod Phys 54:235–268
88. Zoia A, Kantor Y, Kardar M (2007) Distribution of first passage times and distances along critical curves. arXiv:0705.1474v1 [cond-mat.stat-mech]

Stochastic Models of Biological Processes

STEVEN S. ANDREWS[1], TUAN DINH[1], ADAM P. ARKIN[1,2]
[1] Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, USA
[2] Department of Bioengineering and Howard Hughes Medical Institute, University of California, Berkeley, USA

Article Outline

Glossary

Brownian dynamics A level of detail in which each molecule is represented by a point-like particle and molecules move in response to diffusion and collisions

Chemical Fokker–Planck equation (CFPE) Master equation for well-mixed systems that corresponds to the chemical Langevin equation

Chemical master equation (CME) Master equation for the probability that the system has specific integer copy numbers for each type of chemical species; it is exact for a well-mixed system

Chemical Langevin equation (CLE) Approximate stochastic differential equation for well-mixed systems which is based on continuous Gaussian statistics

Direct method An implementation of the Gillespie algorithm

Extrinsic noise In genetic noise studies, expression fluctuations of a gene that arise from upstream genes or global fluctuations

First-reaction method An implementation of the Gillespie algorithm

Gillespie algorithm Exact algorithm for simulating individual trajectories of the CME

Hybrid algorithms Algorithms that are designed to efficiently simulate systems that have multiple timescales

Individual-based spatial models Models that track individual molecules as they diffuse or react

Intrinsic noise Expression fluctuations of a gene that arise from that particular gene

Jump process A process in which the system abruptly changes from one state to another

Optimized direct method A computationally efficient implementation of the Gillespie algorithm

Population-based spatial models Models that track how many molecules of each chemical species are in various spatial compartments

Reaction channel A possible reaction between specified reactant and product chemical species (the terminology distinguishes this meaning from an individual reaction event between single molecules)

Reaction-diffusion equation Deterministic partial differential equation that combines mass action reaction kinetics and normal chemical diffusion

Reaction-diffusion master equation (RDME) Chemical master equation that accounts for diffusion as well as reactions

Reaction rate equation (RRE) Deterministic ordinary differential equation for the net production rate of each chemical species from chemical reactions

Spatial chemical Langevin equation Chemical Langevin equation that accounts for diffusion as well as reactions

Stochastic simulation algorithm (SSA) Alternative term for the Gillespie algorithm

Stoichiometric matrix (ν) Matrix that gives the net production of each chemical species, for each chemical reaction

Tau-leaping method Approximate simulation method for well-mixed systems in which molecule numbers are updated using discrete Poisson statistics

Well-mixed hypothesis Assumption that mixing processes occur faster than the relevant reaction processes

Definition of the Subject

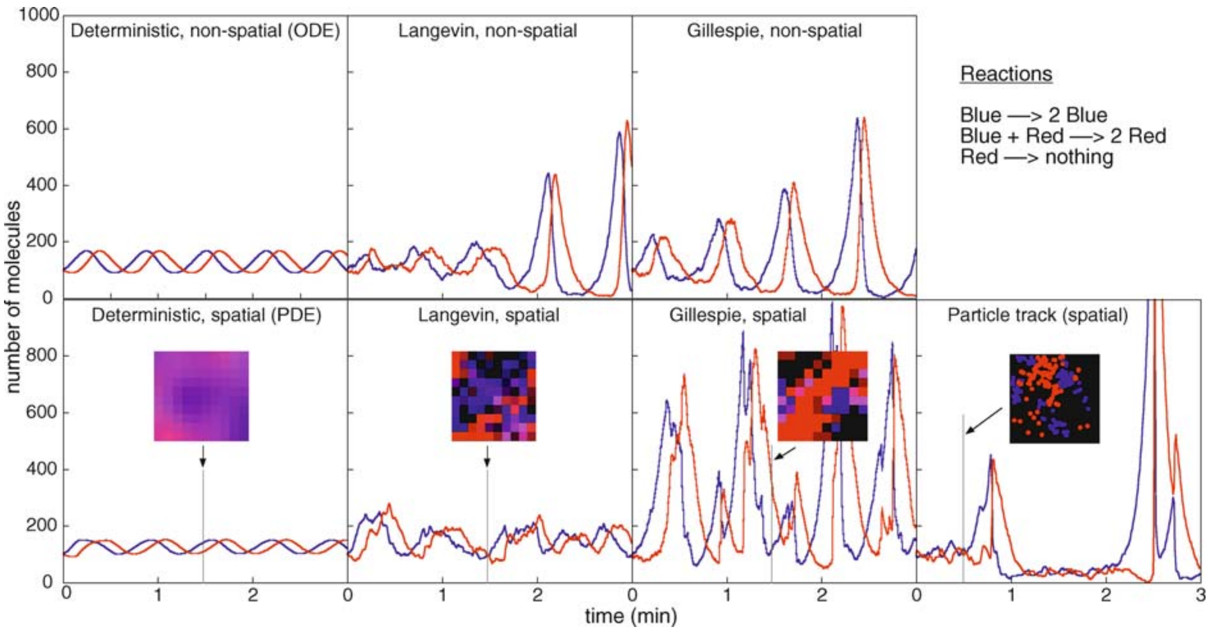
Many processes in cell biology, such as those that carry out metabolism, the cell cycle, and various types of signaling, are comprised of biochemical reaction networks. It has proven useful to study these networks using computer simulations because they allow us to quantitatively investigate hypotheses about the networks. Deterministic simulations are sufficient to predict average behaviors at the population level, but they cannot address questions about noise, random switching between stable states of the system, or the behaviors of systems with very few molecules of key species. These topics are investigated with stochastic simulations. In this article, we review the dominant types of stochastic simulation methods that are used to investigate biochemical reaction networks, as well as some of the results that have been found with them. As new biological experiments continue to reveal more detail about biological systems, and as computers continue to become more powerful, researchers will increasingly turn to simulation methods that can address stochastic and spatial details.

Introduction

Random events are ubiquitous throughout biology. Diffusion, chemical reactions, gene expression, homologous recombination, and most other fundamental biological processes are governed to a large extent by the inherently discrete and stochastic interactions of molecules [1]. In many cases, the random events that occur on very small length and time scales become averaged out when one focuses on larger length or time scales. However, there also exist many examples in which stochastic fluctuations at small scales propagate up to and then influence the system behavior at large scales. Examples range from the swimming trajectories of individual bacteria all the way up to the genetic diversity on which evolution depends.

Research on stochasticity in biochemical systems has received a great deal of attention lately, leading to many recent reviews [2,3,4,5,6,7,8,9]. One reason for its popularity is that the basic designs of many biochemical systems, such as metabolism, cell division, and chemotaxis, are becoming reasonably well understood. Starting from this understanding, researchers are delving deeper to examine the quantitative behaviors of these systems, including the roles of stochastic influences. Also, there is an increasing awareness of the importance of stochasticity in biological systems. For example, it has become clear that noisy gene expression is the rule rather than the exception [9]; this leads to important questions about non-genetic individuality and about biological robustness to gene expression noise. Thirdly, stochasticity is often investigated using computationally demanding simulations. Cheaper and faster computers, as well as improved simulation algorithms, are making it feasible for researchers to investigate more complex biochemical systems, at increasingly realistic levels of detail. Finally, and perhaps most importantly, the last ten years have witnessed incredible progress in experimental biochemical methods, some of which allow direct measurements of stochasticity on microscopic size scales. These methods include gene expression measurements [10,11], flow cytometry [12,13,14,15,16], single molecule detection methods [17,18,19,20], and applications of synthetic biology [21].

At the most fundamental level, the quantum mechanics that describe the dynamics of all physical systems are well-understood and completely deterministic [22]. It is



Stochastic Models of Biological Processes, Figure 1

Simulation results for a simple chemical oscillator using different simulation methods. The Lotka–Volterra system is shown, which shares key features with cellular oscillators such as circadian rhythms. *Insets* show the spatial distributions of molecules at the indicated times. In the *top panels*, note that stochasticity allows the system to drift to large amplitude oscillations and that the Langevin and Gillespie methods yield similar results. In the *bottom panels*, all of which were started with nearly homogeneous initial states, differences arise from the approximations: the PDE simulation has predictable oscillations due to the minimal stochasticity (which is only in the initial state); the Gillespie simulation has larger peaks than the Langevin one because it only allows integer numbers of molecules in each bin; and the particle tracking simulation shows larger and fewer bursts than does the Gillespie simulation because it accurately treats diffusion at all length scales (this difference was reduced with a spatial Gillespie simulation that used smaller subvolumes). Parameters: rate constants are 10 min^{-1}, 8000 nm^3 molec^{-1} min^{-1}, and 10 min^{-1}, for the respective reactions shown in the *top-right corner*, systems start with 100 of each *blue* and *red* molecules, their diffusion coefficients are 100 nm^2 min^{-1}, the volume is 100 nm high and wide by 10 nm deep, and the first three spatial simulations divide this volume into cubic subvolumes that are 10 nm on a side. This figure is reproduced from [8]

only when systems are observed that there arises unavoidable randomness, although these aspects of quantum mechanics remain murky and as close to philosophy as science. More importantly, essentially any system that is governed by nonlinear dynamics, including nearly all physical systems, rapidly becomes chaotic as the system size is increased beyond a few molecules, and thus becomes effectively unpredictable [23]. For all intents and purposes, the diffusive trajectories of individual molecules, and the probabilities of chemical reactions occurring between neighboring reactants at specific times, are fundamentally stochastic.

In the laboratory, partly by design, this stochasticity usually averages out. For systems comprised of many particles, diffusion is observed to be described well by Fick's law of diffusion and chemical reaction kinetics are described well by the deterministic reaction rate equations [24]. The situation is often different within biological cells for several reasons: (i) the program for cellular behavior, the genome, is present at low copy number and yet each gene governs the expression of possibly thousands of proteins, (ii) the low copy numbers of many proteins and mRNA transcripts within cells make random variation of their numbers a relatively large fraction of the total, (iii) stochasticity is often amplified during sequential biochemical steps, of which an especially important case is the sequence of DNA transcription followed by mRNA translation, and (iv) because of spatial organization within cells, it often takes very few proteins in specific locations to achieve large effects on the entire cell dynamics.

This review provides a broad account of stochastic modeling of biological processes. The emphasis is placed on stochastic processes at the cellular level although much of the work presented here also applies to other scales and systems. Our goal is to briefly familiarize the reader with the mathematical forms of the most important equations, the tools for analyzing and simulating them, and some applications for which they have been particularly successful. As shown below, the mathematics, the software implementation, and even the applications of modeling methods are often closely linked.

There are several ways to categorize stochastic biochemical modeling methods. A key designation is whether a method is spatial or non-spatial: spatial models treat spatial organization of proteins and membranes explicitly, whereas non-spatial models include an implicit assumption that mixing processes occur faster than the relevant reaction processes, which is called the *well-mixed hypothesis*. No single modeling method can efficiently capture stochastic dynamics over wide ranges of time scales, so separate methods have been developed that operate at levels of temporal detail that range from nanoseconds to hours. Hybrid simulators combine methods that operate at different timescales to allow the efficient simulation of systems that include both fast and slow processes. Whereas many modeling methods are designed solely to address the reactions in a biochemical reaction network, others also consider system boundaries, mechanics, or the multiple states that proteins can be in. Here, we present non-spatial modeling methods first, followed by spatial methods, with diversions along the way to touch on as many of the other topics as possible. Simulation results from many of the methods that are discussed are compared in Fig. 1, where it is seen that the differences can be quite significant.

Non-Spatial Stochastic Modeling

Deterministic Modeling and Notation

Before focusing on stochastic modeling, it is helpful to introduce the notation and some terminology by summarizing a few aspects of deterministic modeling. As applied to chemical reaction networks, deterministic modeling is based upon ordinary differential equations (ODEs) for the individual chemical reactions. Consider the generic elementary reversible reaction

$$A + B \underset{k_\mathrm{r}}{\overset{k_\mathrm{f}}{\rightleftharpoons}} C\,, \tag{1}$$

where k_f and k_r are the forward and reverse reaction rates, respectively. We assume that the system is kept well-mixed so that diffusion effects can be ignored. The reaction rate equations for components A, B, and C are the ODEs

$$\frac{\mathrm{d}[A]}{\mathrm{d}t} = \frac{\mathrm{d}[B]}{\mathrm{d}t} = -k_\mathrm{f}[A][B] + k_\mathrm{r}[C]\,, \tag{2a}$$

$$\frac{\mathrm{d}[C]}{\mathrm{d}t} = k_\mathrm{f}[A][B] - k_\mathrm{r}[C]\,. \tag{2b}$$

More complex reaction networks are expressed analogously, with one equation for each chemical species and with terms in the equations that represent chemical reactions. From these equations, the reactions can be simulated to show how the concentrations change over time. Or, after setting the left sides of the equations to zero, they can be solved to yield the steady-state chemical concentrations. One can also investigate the dynamic or steady-state behaviors as the reaction rate parameters (k_f and k_r), or initial concentrations, are varied [25]; this can yield phase diagrams for the reactions and additional insight.

It is helpful to generalize the rate equations given above to make them more convenient for computational or analytical work and to show their forms more clearly. First, each chemical concentration is replaced by the variable $Z_i(t)$, where i is an index for A, B, or C and the time-dependence is written out explicitly. Next, the product terms in the equations are replaced by the functions $\tilde{a}_\mathrm{f}(\mathbf{Z}(t))$ and $\tilde{a}_\mathrm{r}(\mathbf{Z}(t))$ for the forward and reverse reactions, respectively; the tildes indicate that molecule quantities are given as concentrations rather than as molecule numbers. These functions are called the *reaction propensities*. Finally, the '+' or '–' signs show the reaction stoichiometry. They are replaced by $\nu_{\mathrm{f}i}$ and $\nu_{\mathrm{r}i}$ for the forward and reverse reactions, respectively, which are elements of the so-called *stoichiometric matrix* (throughout this review, we follow Gillespie's notation [6]). In this example, one unit of each A and B are lost in a unit amount of forward reaction ($\nu_{\mathrm{f}A} = \nu_{\mathrm{f}B} = -1$) as one unit of C is formed ($\nu_{\mathrm{f}C} = 1$); the $\nu_{\mathrm{r}i}$ values have the opposite signs. With these substitutions, the rate equations become

$$\frac{\mathrm{d}Z_i(t)}{\mathrm{d}t} = \nu_{\mathrm{f}i}\,\tilde{a}_\mathrm{f}(\mathbf{Z}(t)) + \nu_{\mathrm{r}i}\,\tilde{a}_\mathrm{r}(\mathbf{Z}(t))\,, \tag{3a}$$

$$\tilde{a}_\mathrm{f}(\mathbf{Z}(t)) = k_\mathrm{f} Z_A Z_B\,, \tag{3b}$$

$$\tilde{a}_\mathrm{r}(\mathbf{Z}(t)) = k_\mathrm{r} Z_C\,. \tag{3c}$$

These equations are trivially extended to arbitrarily large reaction networks. Consider a system with N chemical species that can react via M different *reaction channels* (a "reaction channel" is simply unambiguous terminology for a reaction between specific reactant and product chemical species). The dynamics of this system are given with the *reaction rate equation* (RRE):

$$\frac{\mathrm{d}Z_i(t)}{\mathrm{d}t} = \sum_{j=1}^{M} \nu_{ji}\,\tilde{a}_j(\mathbf{Z}(t))\,. \tag{4}$$

The reaction propensity equations are typically the products of reaction rate constants and the appropriate chemical concentrations, as shown above (Eq. (3b) and (3c)), but they may also describe non-elementary processes such as Michaelis–Menten kinetics. It is worth noting that the state of the system, at any point in time, is fully expressed with the vector $\mathbf{Z}(t)$. This means that the entire trajectory of the system can be deterministically calculated from the RRE and any single $\mathbf{Z}(t)$ snapshot.

The RRE is at the heart of many branches of quantitative biochemistry and systems biology. A great deal of metabolic theory is based on either the steady-state solutions of the RRE, or the set of steady-state solutions that are possible, given only knowledge of the stoichiometric matrix [26,27,28,29]. Studies of biochemical oscillations [25], including the cell cycle [30,31], circadian rhythms [32,33], and certain spatial patterns [34,35] are usually based on the dynamics of the deterministic RRE. Research on biochemical switches, as are found in prion diseases [36], developmental processes [37], and some protein kinase cascades [38], often focuses on the multiple steady-state solutions of the RRE [39]. Deterministic modeling has been, and still is, the conventional modeling method for most biological systems.

The Chemical Master Equation

Although the RRE is tremendously useful, it cannot address the stochastic processes that are inherent to biochemical systems. This is because the RRE arises from a series of approximations to a more physically rigorous stochastic model of chemical reactions [40,41].

As above, we assume that diffusive processes are much faster than reactive ones [8], which allows us to ignore spatial organization (this assumption is usually valid for genetic and metabolic networks, but often invalid for signaling networks). Nevertheless, the stochasticity of diffusion plays an essential role because it makes the precise timing of individual reactions effectively random. These reactions occur in abrupt transitions, in which reactants are converted effectively instantaneously into products, making this a type of *jump process*. Also, random diffusion causes the system to rapidly lose any memory of its prior states, and thus of the sequence of reactions that led up to the current state. This independence of the system dynamics on its history, called the Markov property, implies that the probability that a specific reaction occurs depends only on the state of the system at that time [41].

Because of the random reaction timing, reactant concentrations do not follow the deterministic trajectory that is predicted by the RRE. Instead, many concentration trajectories are possible, of which a single effectively random one actually occurs. There are two primary ways to investigate the possible trajectories with computational methods. One can simultaneously track the probability of every possible outcome or one can simulate many independent stochastic trajectories and then analyze them as one would with several repetitions of an experiment. These methods are described in this and subsequent sections, respectively.

To mathematically track the probability that the system is in each possible state, it is helpful to first replace the vector of chemical concentrations that was introduced earlier, $\mathbf{Z}(t)$, with a vector of integer-valued molecule numbers, $\mathbf{X}(t)$. These are related to each other, within round-off error, through the volume of the system, which is given as Ω,

$$\mathbf{X}(t) \simeq \Omega \mathbf{Z}(t) \,. \tag{5}$$

The state of the stochastic system is fully captured by $\mathbf{X}(t)$. The probability that the system is in state $\mathbf{x}$ at time t, given that it started in state $\mathbf{x}_0$ at time t_0, is written as $P(\mathbf{x}, t \mid \mathbf{x}_0, t_0)$. This probability changes over time because chemical reactions can transfer the system either into this state from other ones, or out of this state and into others. These possible transitions are combined to yield the *chemical master equation* (CME) [6,42]:

$$\frac{\partial P(\mathbf{x}, t \mid \mathbf{x}_0, t_0)}{\partial t} = \sum_{j=1}^{M} \Big[a_j(\mathbf{x} - \underline{\boldsymbol{\nu}}_j)\, P(\mathbf{x} - \boldsymbol{\nu}_j, t \mid \mathbf{x}_0, t_0) - a_j(\mathbf{x})\, P(\mathbf{x}, t \mid \mathbf{x}_0, t_0) \Big] \,. \tag{6}$$

The sum is over the reaction channels that can occur in the system. The two terms within brackets give the rate at which the probability of being in state $\mathbf{x}$ increases or decreases over time because of reactions into or out of state $\mathbf{x}$, respectively. These are proportional to the reaction propensities for the respective reactions. They are also proportional to the probability that the system was in the starting state, because the system can only leave a state if it was there in the first place.

The reaction propensity $a_j(\mathbf{x})$, given here without a tilde because it is for molecule numbers rather than concentrations, is a probability density: $a_j(\mathbf{x})\mathrm{d}t$ is the probability that exactly one reaction of type j will occur in a system in state $\mathbf{x}$ within the next $\mathrm{d}t$ amount of time. This microscopic propensity function is as central to stochastic chemical kinetics as its macroscopic analog is to the reaction rate equation. However, the microscopic propensity rests on a solid microphysical basis, and has in fact been shown to have an exact solution for a well-stirred thermally-equilibrated gas-phase system [43]. For such a system, the propensity function is

$$a_j = h_j c_j \,, \tag{7}$$

where h_j is the number of distinct combinations of individual reactants for reaction j and c_j is the probability density for one of those reactions to occur. That is, $c_j(t)\mathrm{d}t$ is the probability that a randomly selected set of reactants for reaction j will collide and react in the next infinitesimal time interval dt. For a variety of reaction mechanisms, the c_j values can be calculated quite accurately from only the physical properties of the system [43].

The primary approximation made for the CME is that the reactive system is well-mixed. This implies both that there is no spatial organization and that there are no significant correlations between successive reactions (the Markov property). Examples of correlated reactions include metabolite channeling, which is the transfer of a metabolite from one enzyme to the next before it has a chance to equilibrate into the cytoplasm [44], and geminate recombinations, which are multiple bindings between molecules that bind reversibly [45]. The well-mixed statement also encompasses the assumption that the system is isothermal, which is typically the case in biological systems.

Because the CME becomes computationally intractable with any but the simplest systems, some recent work has focused on efficient solution methods. An algorithm called the finite state projection method accomplishes this by projecting a matrix form of the CME onto a smaller space [46,47]. By choosing the size of the projection space, the accuracy can be adapted to any level of precision. Less formal methods for state-space reduction of the CME have been proposed as well [48]. Another method uses a sparse grid, which can work efficiently for up to 10 proteins [49]. Work has also gone into separations of the CME into fast and slow components, as described in the section on hybrid methods [42,50,51].

Applications of the Chemical Master Equation

The solution of the CME suffers from the so-called "curse of dimensionality" as the size of the state space, and hence the number of equations, increases exponentially with the number of chemical species involved. Except for very small and simple systems, it is extremely difficult to obtain solutions of the CME, either analytically or numerically. However, a few papers do report quite interesting results from direct simulations of the CME.

The master equation was used to investigate the dynamics of transiently denatured segments of double stranded DNA [52]. The authors derived the dynamics, formation rates, and lifetimes of these "bubbles", which can be compared to fluorescence correlation microscopy experiments of fluorescently tagged base pairs [53]. In another use of the CME, studies on molecular motors [54,55] demonstrate how the load-velocity curve, including rectified motion, arises from nucleotide triphosphate binding free energies. These works more fully investigate ideas on thermal ratchets that were presented previously [56]. A particularly intriguing study on the copy number control system for bacterial plasmids [57] showed that stochasticity in a regulatory portion of a system can actually decrease the stochasticity elsewhere in the system. This "stochastic-focusing" changes the behaviors of gradual-response systems towards those of threshold systems [58,59] in a manner that is analogous to the oscillation enhancement that stochastic resonance can create in oscillating systems [60]. These results contradict the widely held belief that an increase in stochasticity in one portion of a system will necessarily increase the stochasticity everywhere downstream of it. Studies of simple signal transduction motifs have shown how the predictions of the RRE can be qualitatively wrong compared to the CME treatment in that the stochastic systems might be bistable or oscillate when the deterministic system has one stable state [40,61]. Many of these studies that used the CME also used other theoretical techniques as well, which allow fruitful comparisons between the methods.

The Gillespie Algorithm

Because of the challenges in working with the CME, it is most often investigated using a Monte Carlo approach in which individual sample trajectories are simulated. These simulations can be exact or approximate. In this context, "exact" means that if the simulation were run many times, the distribution of simulated trajectories would agree exactly with that which would be predicted by an analytical solution, were it obtainable, of the chemical master equation. Exactness implies nothing about the validity of the CME or about the limitations of the computational accuracy (such as round-off errors and imperfect pseudo-random number generators), but only that no further approximations are made beyond those that are assumed by the CME.

In 1976, Gillespie introduced an exact algorithm for simulating the CME [6,62,63] which is called the *stochastic simulation algorithm* (SSA) in his papers, but is better known as the *Gillespie algorithm*. This algorithm cycles through three portions: (i) generate the time step until the next reaction, (ii) determine which reaction that will be, and (iii) execute the reaction by advancing the time and molecule counts to reflect it. The Gillespie algorithm was introduced with two varieties, called the *direct method* and the *first-reaction method*. In the former, the time step to the next reaction, τ, and the reaction number, j, are chosen from the following probability distributions:

$$P(\tau) = a\mathrm{e}^{-a\tau}\,, \tag{8a}$$

$$P(j) = \frac{a_j(\mathbf{X})}{a}\,. \tag{8b}$$

The variable a represents the summed reaction propensity,

$$a \equiv \sum_j a_j(\mathbf{X})\,. \tag{8c}$$

In the first-reaction method, a time step, τ_j, is generated for each possible reaction channel. Again, these are exponentially distributed random numbers,

$$\tau_j = a_j(\mathbf{X})\,\mathrm{e}^{-a_j(\mathbf{X})\tau}\,. \tag{9}$$

The smallest of these time steps is chosen as the next simulation time step, while its subscript dictates the reaction channel that is executed at that time. The direct method is usually preferred because it is a little easier to program and runs slightly faster with a simple implementation. However, the latter has been favored as a basis for improvements on computational efficiency.

The exactness of the Gillespie algorithm comes at the cost of its being computationally demanding. Even if one simulation can be performed reasonably quickly, such as in a few minutes, this can still be too slow to investigate the behaviors of hundreds of mutant cells or to explore different regions of parameter space. Thus, much effort has been devoted to improving the efficiency of the Gillespie algorithm, while still maintaining exactness. These methods are all based upon either the direct method or the first-reaction method, but are carefully designed, usually with priority queues or other indexing methods, so that internal variables are recalculated as infrequently as possible [6,64,65,66,67]. Of these, it appears that the *optimized direct method* is probably best for most practical problems [66]. Because the faster methods are significantly more difficult to program than the original ones, both sets of methods are still commonly used.

The computational intensity of the Gillespie algorithm, even with more efficient implementations, makes it difficult to perform sensitivity analyzes. In these analyzes, one investigates the extent to which the results depend upon input parameters, which can helpful for determining which parameters need additional experimental investigation or which are particularly important for system control. An algorithm for stochastic sensitivity analysis was recently developed and applied to biochemical reaction networks [68]. It involves the addition of just two steps to the basic loop of the Gillespie algorithm.

Another difficulty of the Gillespie algorithm, which also applies to simulations of the RRE and other algorithms presented below, is called *combinatorial explosion*. Suppose a scaffold protein has several sites with which it can bind other proteins, and suppose that each of those proteins can bind to none, one, or two phosphate groups (this is the situation for the Ste5 protein in the yeast pheromone response pathway [69]). There are clearly a tremendous number of possible binding states that the scaffold protein can be in, each of which has to be treated as a separate chemical species. Just listing all of these states is tedious, and simulating their reaction dynamics with the Gillespie algorithm is very slow. One solution is to not list every possibility when the simulation starts, but to create states when they are needed and to destroy them when they are no longer required [65]. Another option is to use an algorithm that is implemented in a program called StochSim [70,71]. Unlike the Gillespie algorithm, this one does not stochastically choose reactions to execute, but it instead chooses reactant pairs from the pool of existing molecules. A probabilistic scheme is used to determine if these reactants should be made to react with each other.

What if the system volume changes as a function of time? This might seem like an unusual concern, but it occurs during cell growth (and cell division) and it affects the reaction propensities. The necessary modifications to the Gillespie algorithm were recently derived, which are likely to be particularly useful for relatively slow processes, such as protein production from infrequently expressed genes [72].

Applications of the Gillespie Algorithm

Models based on the Gillespie algorithm have provided critical insights into the stochastic nature of gene expression [3,73,74]. In particular, fluctuations in the rates of gene transcription are amplified at the translation stage to yield highly erratic time patterns of protein production [75]. When multiple regulatory proteins act together, or compete with each other, this randomness is amplified further because of the random sequence of protein bursts [75]. These effects were shown to stochastically switch a model of phage-λ between the bistable lysis and lysogeny states [76], with results that are consistent with experimental ones. Stochastic gene expression is also used by many pathogenic organisms to randomly switch their surface features so they can evade host responses [77], may be used by the HIV virus to stochastically delay viral expression long enough for transformation of its activated T-cell host to a memory cell and thereby trap HIV as a latent phage [16], can establish asymmetries that determine cell differentiation [78], and can cause circadian clocks to lose synchrony [21,79]. In fluctuating environments, stochastic gene expression can permit an isogenic bacterial population to grow faster than it would if all individuals were phenotypically homogeneous [77,80,81].

From combined modeling and experimental approaches, the dominant noise sources in the stochastic expression of a specific gene are: (i) the expression fluctuations of that particular gene, which is called the *intrinsic noise*, (ii) noise that is transmitted to it from upstream genes, and (iii) global noise that affects all genes. The latter two sources are often combined and called *extrinsic noise*, which is the total noise source that is extrinsic to that specific gene [12,74,82,83,84]. Noise arises at both the transcription and translation stages, for which the relative importance depends on the strength of the promoter and on whether prokaryotic or eukaryotic transcriptional machinery is used [14,15,75,85,86]. Direct measurements of gene expression have generally confirmed the predictions made by stochastic simulations [11,12,13,15,17,18,21,87].

Gene expression appears to be unavoidably stochastic, and this randomness is usually amplified at each stage, so how does biology function reliably amidst all the noise? This is a central topic of many papers on biological robustness [5,9,88,89,90,91], several of which use the Gillespie algorithm or other types of stochastic modeling. One answer is that many reaction network structures are inherently less susceptible to noise than others. These include ones in which the reaction rates do not depend on the number of mRNA transcripts [92], certain scale-free reaction networks [93], and networks that are designed to function near saturation [94] (analogous to binary logic). Secondly, there are several mechanisms for biological robustness to noise, including negative feedback, integral feedback [95], checkpoints, and redundancy [9]. Because gene expression noise is usually detrimental to biological function, it has been suggested that there is active selection for robustness mechanisms [96,97].

Approximate Stochastic Methods

Because every reaction is simulated individually in the Gillespie algorithm, it is unavoidably computationally demanding, even with the algorithmic methods that have been developed to speed it up. To address this, several approximate methods have been developed.

The most accurate of these approximate methods is called the *tau-leaping* method [6,98,99,100]. In contrast to the Gillespie algorithm, the τ-leaping method uses a simulation time step which is long enough that many individual reactions are likely to occur during the time interval. The reaction propensities ought to change slightly as each reaction occurs to reflect the new chemical populations, although this algorithm uses the assumption that changes within a single time step are negligible. This is the sole approximation made for the τ-leaping method. Each reaction is considered to be an independent event (a consequence of both the well-mixed hypothesis and the constant reaction propensities), so the number of reactions that occur during time step τ for reaction channel j is a Poisson-distributed random variable; it is denoted k_j and has a mean value equal to the reaction propensity $a_j(\mathbf{X}(t))$. The formula used by τ-leaping that updates the system state over one time step is

$$\mathbf{X}(t+\tau) = \mathbf{X}(t) + \sum_{j=1}^{M} k_j \boldsymbol{\nu}_j \,. \tag{10}$$

The algorithm alternates steps in which the state of the system is updated and those in which new reaction propensities are calculated.

As the time step is reduced to zero, the τ-leaping simulation method approaches that of the Gillespie algorithm, although with more computational overhead. In the other direction, increasing the time step makes the simulation becomes more and more approximate. It has been suggested that τ should be chosen by first predicting the change in molecular populations over time using deterministic methods; then, the time step is chosen so that no molecular species is likely to change its population during this time step by more than some pre-determined fraction of its total population [98]. A difficulty that can occur with τ-leaping (which is also an issue with ODE integration), is that it is possible for a molecular species to be assigned a negative population. Several methods have been proposed to avoid this problem, some of which are also able to improve the performance of the algorithm in other ways as well [101,102,103]. Despite several papers on the development of τ-leaping, this method has yet to be applied to novel biological problems.

Two additional approximations allow the application of many more theoretical methods. First, the vector that defines the state of the system, $\mathbf{X}(t)$, is allowed to take on real values as well as integer values. As one would expect, this is usually a reasonable assumption for large chemical populations and a poor assumption for low copy numbers. In particular, it can a very poor approximation in cases where there might become no copies of a chemical species at all; an approximate value of, say, 0.01 protein copies can lead a system to entirely different outcomes than one would find with exactly 0 copies. The second approximation is to replace the Poisson-distributed random variables that were used in the τ-leaping algorithm with Gaussian-distributed random variables [104]. The effect of this change again decreases as the copy numbers of chemical species are increased. It also decreases as longer time steps are used because that leads to more individual chemical reactions per time step. These approximations allow the updating equation of the τ-leaping algorithm to be replaced by a stochastic differential equation called the *chemical Langevin equation* [105,106,107] (CLE),

$$\frac{\mathrm{d}X_i(t)}{\mathrm{d}t} = \sum_{j=1}^{M} \nu_{ji}\, a_j(\mathbf{X}(t)) + \sum_{j=1}^{M} \Gamma_j(t)\, \nu_{ji} \sqrt{a_j(\mathbf{X}(t))}\,. \quad (11)$$

The first term is simply the reaction rate equation that is given above (Eq. (4)), but for molecule counts rather than concentrations. The second term adds Gaussian noise to the deterministic result, where $\Gamma_j(t)$ represents a temporally uncorrelated, statistically independent Gaussian white noise with mean 0 and variance 1. In other words, the integral of $\Gamma_j(t)$ is a one-dimensional continuous random walk. The chemical Langevin equation describes a continuous Markov process which is an approximation of the jump Markov process that underlies the chemical master equation. Simulations with the CLE, which are based on an equation that is quite similar to Eq. (11) [106], yield single stochastic trajectories of the system state, much like simulations with the Gillespie algorithm or the τ-leaping method.

Alternatively, instead of following a single system as it moves along one of its many possible stochastic trajectories, it is possible to focus on a single portion of the state space to see how likely it is that the system will be in this region of state space as a function of time. This latter picture is described by *chemical Fokker–Planck equation* [41,105,107] (CFPE),

$$\begin{aligned}\frac{\partial P(\mathbf{x},t \mid \mathbf{x}_0,t_0)}{\partial t} &= \\ &-\sum_{i=1}^{N} \frac{\partial}{\partial x_i}\left[\left(\sum_{j=1}^{M} \nu_{ji} a_j(\mathbf{x})\right) P(\mathbf{x},t \mid \mathbf{x}_0,t_0)\right] \\ &+ \frac{1}{2}\sum_{i=1}^{N} \frac{\partial^2}{\partial x_i^2}\left[\left(\sum_{j=1}^{M} \nu_{ji}^2 a_j(\mathbf{x})\right) P(\mathbf{x},t \mid \mathbf{x}_0,t_0)\right] \\ &+ \sum_{\substack{i,i'=1 \\ i<i'}}^{N} \frac{\partial^2}{\partial x_i \partial x_{i'}}\left[\left(\sum_{j=1}^{M} \nu_{ji}\nu_{ji'} a_j(\mathbf{x})\right) P(\mathbf{x},t \mid \mathbf{x}_0,t_0)\right]. \end{aligned} \quad (12)$$

The first term, often called the drift term, arises from the deterministic behavior of the system. The latter two terms, collectively called the diffusion term, represent the stochastic deviations away from deterministic behavior. Because the CFPE represents continuous processes, it is significantly more analytically tractable than the chemical master equation.

To compute the probabilities of possible system behaviors using the CFPE, state space is usually discretized into a grid and then the CFPE is integrated using standard numerical methods [108,109]. This analysis method is similar to that employed for the CME, but is usually less computationally demanding because the discretized state space is typically significantly coarser. Nevertheless, the dimensionality of state space still increases exponentially with additional chemical species, so the CFPE still suffers from the curse of dimensionality.

Applications of Approximate Stochastic Methods

Perhaps because stochastic simulations are still a relatively new field of study, many studies with the CLE and CFPE focus more on the mathematical techniques than on the biological applications [109,110,111,112,113]. The approximate CLE and CFPE have been shown to yield results that are in good agreement with exact simulations for a reversible isomerization reaction, even with very few molecules [107].

The CLE was used to analytically investigate the role of noise-induced phenomena in enzymatic futile cycles, which is a motif that is common to many biochemical networks [61]. The analysis indicated that the presence of external noise is sufficient to induce switching bistability in the system, a phenomenon that is often attributed to feedback loops [25]. In combination with experimental data, the CLE was also used to show that translational efficiency is the predominant source of intracellular noise for a single-gene system [15]. The Fokker–Planck equation has been used to model cell growth [111,112] and cell migration [113,114]. Of particular interest, the Fokker–Planck equation has provided a convenient framework to describe the behaviors of molecular motors [109]. A motor protein is approximated as a diffusion particle in a periodic asymmetric free-energy surface. Under the input of chemical energy, the motor switches stochastically between different potentials that describe distinct biochemical states of the motor. The model has been used to explain key experimental observations for molecular motors, most notably for the F_1F_0-ATPase system [115] and a bacterial flagellar motor [109,116].

Hybrid Algorithms

Systems that involve multiple time scales provide major simulation challenges. If the fast time scale is simulated with high precision, then the simulation takes too long for the dynamics of the slow one to be observed with any reasonable efficiency. On the other hand, if the simulation time steps are optimized for the slow timescale, then they are too long for the fast reactions and numerical errors become problematic. In the language of differential equations, these are stiff systems which require special solution techniques. For stochastic simulations with multiple timescales, several methods have been developed recently.

One class of hybrid methods focuses on new mathematics to allow approximations of the Gillespie algorithm, or related algorithms, to function with reasonable accuracy over a wide range of timescales [42,50,51,117,118,119,120,121]. The other class generally involves the coupling of multiple simulators, usually including ODE, Langevin, and/or Gillespie; the high-population molecular species are simulated with less stochastic detail and the low-population species are simulated with more stochastic detail [122,123].

Spatial Stochastic Modeling

Most biological systems are highly organized. For example, *Escherichia coli* bacteria have helical cytoskeletons, polar-localized proteins, centrally positioned chromosomes, and elaborate flagellar motor complexes. Eukaryotes are even more organized, with elaborate organelles, microtubules and other complex cytoskeletal elements, motor proteins that shuttle back and forth, and carefully controlled cell shapes. Even phages display remarkable order in the way the DNA is packed into the outer shell. Where does this order come from? And how does this order influence the biochemical reaction network? These questions are being investigated with new imaging experiments [124,212] and with new computer simulation methods that can account for spatial heterogeneity. These spatial simulation methods are the focus of this section.

Spatial simulations have been used to investigate a wide variety of topics. These include: morphogen gradients across *Drosophila* and *Xenopus* oocytes [125,126,127], the *Escherichia coli* cell division plane localization system [128,129,130,131,132,133,134] (see Sect. "Box 1: The *E. Coli* Min System"), intracellular signaling [135,136,137,138], and rebinding of ligands to receptor complexes [139,140,141].

As described above, many successful biochemical models do not account for spatial heterogeneity; in fact, non-spatial models are in the vast majority. Typically, non-spatial models get away with ignoring space because they model dynamics that occur more slowly than the time it takes for a molecule to diffuse across a cell, because they investigate processes that are not intrinsically spatial, and because they do not demand high quantitative accuracy. As the tools are becoming available, including both fast computers and new software algorithms [45,142,143,144], the interest in including spatial detail is increasing. These spatial models can be either deterministic or stochastic, of which our primary focus is on the latter ones.

As with the non-spatial methods that are described above, stochastic effects in spatial models arise from the discreteness of molecules. This leads to fluctuations in the numbers of molecules, which are typically on the order of the square root of the number of molecules in the appropriate characteristic volume (near steady-state and equilibrium points, but frequently greater near critical points). In spatial models, the characteristic length scale is no longer the size of the entire system but is dictated by the length scale of the spatial heterogeneity. With the shorter length scale, the characteristic volume size is reduced, fewer molecules are in these volumes, and stochastic effects increase. Thus, stochastic simulations can be required for spatial models, even if they were not needed for the corresponding non-spatial model. There are also other good reasons to model stochastic effects in spatial simulations. Many spatial phenomena, such as noise that arises from ligand rebinding [141], cannot be adequately treated without considering the detailed molecular interactions. Finally, a model is only as good as its weakest aspect. If one increases the accuracy of a model in one way, such as by accurately treating either space or stochastics, then the benefits may not be realized until the other aspect is addressed as well.

Schemes for investigating a chemical system with spatial and stochastic detail can be classified by whether they consider molecules within populations or as individuals. In the former case, space is divided it into small subvolumes, whereas in the latter, space is continuous. These classes are described in detail below. Another approach is lattice-based methods [145,146,147,148,149]. However, we do not discuss them here because they are rarely used for quantitative modeling. Furthermore, the underlying lattice geometry usually affects the results, thus making them less realistic.

Population-Based Spatial Models

In a top-down approach towards spatial modeling, one starts with a simple, deterministic, macroscopic description and then adds successive layers of detail. In this case, the natural starting point is with the standard textbook descriptions of chemical reactions and diffusion [24]. Reactions are described with mass action reaction kinetics expressed with the reaction rate equation that was discussed above (Eq. (4)). Diffusion is described with the diffusion equation [24], also called Fick's second law of diffusion, which is

$$\frac{\partial Z_i(\mathbf{r},t)}{\partial t} = D_i \nabla^2 Z_i(\mathbf{r},t)\,. \tag{13}$$

In an extension of the definition given before, $Z_i(\mathbf{r},t)$ is the concentration of component i at the 3-dimensional position $\mathbf{r}$ and time t. D_i is the diffusion coefficient for component i.

Because reactions and diffusion occur simultaneously, the respective equations are combined to express the simultaneous effects of both processes to yield the *reaction-diffusion equation*,

$$\frac{\partial Z_i(\mathbf{r},t)}{\partial t} = \sum_{j=1}^{M} \nu_{ji}\,\tilde{a}_j(\mathbf{Z}(\mathbf{r},t)) + D_i \nabla^2 Z_i(\mathbf{r},t)\,. \tag{14}$$

This partial differential equation (PDE) underlies a great deal of theory on chemical and biological pattern formation [126,136,150,151]. The Virtual Cell computer program [152] is general-purpose software that simulates the reaction-diffusion equation. It has been used primarily to explore spatial effects in intracellular signaling [153,154,155].

The reaction-diffusion equation is deterministic, so it captures neither the discreteness of reaction events nor the Brownian motion processes that underlie diffusion. It is possible to add this stochasticity directly into the deterministic theory but that would create a set of coupled stochastic scalar field equations, which would be extraordinarily complicated. Neither the deterministic nor the stochastic PDEs are tractable to work with analytically for any but the very simplest systems except, perhaps, in steady-state. Thus, most analysis is either computational or approximate.

In most such analyses, the equations are first simplified by dividing the system volume into an array of small cubic subvolumes, each with width l. This spatial discretization changes the diffusion portion of the reaction-diffusion equation into a discrete form:

$$\frac{\mathrm{d}Z_{i,k}(t)}{\mathrm{d}t} = \sum_{j=1}^{M} \nu_{ji}\,\tilde{a}_j(\mathbf{Z}_k(t)) + \frac{D_i}{l^2}\sum_{k'}\left[Z_{i,k'}(t) - Z_{i,k}(t)\right]\,. \tag{15}$$

The index k denotes the subvolume number, much as $\mathbf{r}$ represented the spatial location. The latter summation in this discrete reaction-diffusion equation extends over all nearest neighbors of subvolume k, denoted by k'. Because the description of space was changed from continuous states to discrete states, much like the discrete kinds of molecules that are labeled by the index i, diffusion is now formally identical to reactions. The "reaction rate constant" for diffusion [156] between one subvolume and its neighbor is D_i/l^2. Because of this mathematical equivalence, much of the following discussion on the stochastic simulation of the reaction-diffusion equation parallels the discussion presented earlier on non-spatial stochastic simulations.

The first spatial stochastic equation that we present is the one that accounts for the least detail. It is the *spatial chemical Langevin equation*, which results from adding white Gaussian noise to the discrete reaction-diffusion equation. It is

$$\begin{aligned}\frac{\mathrm{d}Z_{i,k}(t)}{\mathrm{d}t} = {} & \sum_{j=1}^{M} \nu_{ji}\left[\tilde{a}_j(\mathbf{Z}_k(t)) + \Gamma_j(t)\sqrt{\tilde{a}_j(\mathbf{Z}_k(t))}\right] \\ & + \sum_{k'}\left\{\frac{D_i}{l^2}\left[Z_{i,k'}(t) - Z_{i,k}(t)\right]\right. \\ & \left. + \Gamma_{k'}(t)\sqrt{\frac{D_i}{l^2}}\left[\sqrt{Z_{i,k'}(t)} - \sqrt{Z_{i,k}(t)}\right]\right\}\,. \end{aligned} \tag{16}$$

In an extension to what was presented before, $\Gamma_j(t)$ and $\Gamma_{k'}(t)$ represent temporally uncorrelated, statistically independent Gaussian white noises [106]. This is a specific example of the more general multivariate Langevin equation; it, and the multivariate Fokker–Planck equation, have been explored in depth [41,105]. However, the more specific spatial chemical Langevin equation has essentially never been used, investigated mathematically, or simulated. The sole exception that we are aware of was its simulation for a figure for a tutorial article [8] (those results are reproduced in Fig. 1).

The spatial chemical Langevin equation captures stochasticity reasonably accurately for systems in which there are many molecules per subvolume but not for those with few molecules per subvolume. Errors arise both because Gaussian white noise is the incorrect fluctuation distribution [100,104] and because it treats molecule amounts as continuously variable quantities. These are addressed by moving to the next level of detail in which the continuous molecular concentrations are replaced by discrete numbers of molecules. This changes the temporally continuous reaction and diffusion processes to stochastic jump processes. The *reaction-diffusion master equation* [156,157,158,159] (RDME) describes the time dependence of the system at this level of description. It is

$$\frac{\mathrm{d}P(\mathbf{x},t)}{\mathrm{d}t} = \sum_{j=1}^{M}\Big[a_j(\mathbf{x}-\underline{\boldsymbol{\nu}}_j)P(\mathbf{x}-\boldsymbol{\nu}_j,t) - a_j(\mathbf{x})P(\mathbf{x},t)\Big] + \sum_{i=1}^{N}\frac{D_i}{l^2}\sum_{k,k'}\Big[(X_{i,k}+1)\,P(\ldots,X_{i,k}+1, X_{i,k'}-1,\ldots,t) - X_{i,k}P(\mathbf{x},t)\Big]\,. \tag{17}$$

$P(\mathbf{x},t)$ is the probability that the system is in state $\mathbf{X}=\mathbf{x}$ at time t, $X_{i,k}$ is the number of molecules of type i in subvolume k, $\mathbf{X}$ is the vector of all $X_{i,k}$ values, and a_j is the propensity of reaction j.

The RDME expresses as much detail as is possible through these successive improvements of the reaction-diffusion equation. It is tempting to think of it as the fundamental equation for reactions and diffusion, and thus the basis for a statistical theory of chemistry. In fact, it is sufficiently accurate for most systems, but it nevertheless involves approximations that can be important in some situations. Firstly, neither of the starting equations, which are mass action kinetics and Fickian diffusion, are completely accurate even for very large systems. Mass action kinetics does not address the increased reaction rates that occur on extremely short time scales, which arise from reduced spatial correlations [160,161]. Nor does it address the geminate recombinations that can occur between the products of a dissociation reaction [162,163]. The diffusion equation is usually quite accurate for dilute solutions but fails for highly crowded ones [164,165], including most biological systems [166]. Secondly, the discretization of space into small subvolumes can also lead to inaccuracies, or exacerbate the inaccuracies just mentioned. The subvolume sizes must not be so small that they impinge on the microscopic details of the reaction or diffusion processes. This means that they need to be significantly larger than single molecules and larger than the mean free path lengths of diffusion [156,167]. Conversely, the subvolumes must not be so large that there would be appreciable concentration gradients across them. This means that the subvolume width needs to be less than the reactant correlation length. The correlation length is hard to predict but is at least as large as the average distance that a reactant travels before it reacts, called the reactive mean free path [156,167].

The RDME is even more intractable than the non-spatial chemical master equation because of the addition of spatial states and the many transitions that can occur between the spatial states. These additional states and transitions also make stochastic simulations of the RDME with Gillespie's direct algorithm extremely slow [157]. Several faster algorithms have been developed to address this problem. The "next reaction method" of Gibson and Bruck [64] was adapted to spatial simulations [168], and then further improved, to yield the "next subvolume method" [132,142]. Also, a fast version of the direct method [169] has been developed for spatial simulations [167]. All of these methods yield exactly the same results as Gillespie's original methods [62,63] but use carefully optimized data structures to minimize the number of computations.

A separate challenge with simulating the RDME concerns the cubical subvolumes into which space was discretized. Biological systems rarely have square corners, so the basic theory requires adaptation to account for realistic boundaries. In one approach, the mathematics was developed for dividing boundary subvolumes into two separate portions [170]. Using another approach, the theory was developed for curved surfaces, which was implemented in the MesoRD program [132,171]. Although it has not been developed yet, it has been proposed that automatic mesh refinement could simultaneously account for complex boundaries and lead to significant computational efficiencies [167].

Along with simulations of the *E. coli* Min system, presented in Sect. "Box 1: The E. Coli Min System", population-based spatial stochastic models have been used for a variety of test systems. In the first implementation of a spatial Gillespie algorithm, Stundzia and Lumsden used a one-dimensional simulation to demonstrate stochastic calcium wave propagation [157]. Elf and Ehrenberg showed that spatial and stochastic effects can cause an intrinsically bistable system to lose its global hysteresis through the formation of spatial domains [142]. In a third study, Isaacson and Peskin demonstrated their method for simulating porous boundaries with a model that includes transcription, translation, and nuclear membrane transport [170].

Individual-Based Spatial Models

In a bottom-up approach to spatial modeling, one starts with a very detailed consideration and then makes successive approximations. A convenient place to start is by considering every individual molecule in the system, along with some of the molecular structures. The motions of these molecules are governed by physical forces including steric repulsion, bond mechanics, and electrostatics. The simulation of the motions that result from these forces is called molecular dynamics [172]. Molecular dynamics can yield very accurate results but is so computationally intensive that it is rarely used for more than hundreds of cubic nanometers of volume or more than tens of nanoseconds of time. These size and time scales are too confining for studying biochemical reaction networks, so approximations are made.

At the Smoluchowski level of detail, all solvent molecules are ignored, solute molecules are treated as spheres, diffusion proceeds stochastically, and molecular rotation, molecular momentum, and long-range intermolecular forces are all ignored. This is a vast simplification, but is often valid. It is usually reasonably accurate for size scales that are larger than a few nanometers and for timescales that are longer than a few nanoseconds, constraints that are acceptable for an enormous range of chemical and biological phenomena.

For diffusion at the Smoluchowski level of detail, the effects of solvent-solute interactions on the solute motion are approximated by assuming that solute molecules diffuse with mathematically ideal Brownian motion [173,174]. This is a key approximation that replaces the deterministic molecular motions that result from solvent collisions with stochastic trajectories. It is often the only source of stochasticity in the theory, or in simulations that derive from this individual-based approach. More precisely, the position of molecule i at time t is given with the probability density $p_i(\mathbf{r}, t)$, which evolves over time according to the master equation

$$\frac{\partial p_i(\mathbf{r}, t)}{\partial t} = D_i \nabla^2 p_i(\mathbf{r}, t) \,. \tag{18}$$

This equation is nearly identical to the diffusion equation, given above (Eq. (13)), differing only in the definitions of the variables and the interpretation. Now, it is not a population of molecules that diffuse, but the positional probability density for a single molecule.

Because it is so simple, the diffusion master equation is analytically tractable, in contrast to the other master equations that were discussed. One result is an entire body of analytical theory on diffusion-influenced reactions [160,175]. Nevertheless, it too becomes unmanageable for systems that have several interacting molecules, so it is simulated with a technique called *Brownian dynamics* [176,177,178,179,180]. In this method, molecules have well-defined point-like positions which are updated at each simulation time step using random displacements. The displacements are chosen by solving the diffusion master equation for molecule i, which is taken to be at the well-defined position $\mathbf{r}_0$ at time t_0. One simulation time step later, at time $t_0 + \Delta t$, the probability density for the molecule's position is found to be a 3-dimensional Gaussian density that is centered at $\mathbf{r}_0$,

$$p_i(r, \Delta t) = \frac{1}{s_i^3 (2\pi)^{3/2}} \exp\left[-\frac{(\mathbf{r}-\mathbf{r}_0)^2}{2 s_i^2}\right] . \tag{19a}$$

The standard deviation of this Gaussian, called the root mean square step length, is

$$s_i = \sqrt{2 D_i \Delta t} \,. \tag{19b}$$

Brownian dynamics simulations provide accuracy that is below that of molecular dynamics, but still captures single molecule behavior.

Brownian dynamics has been used extensively for examining the rates of diffusion-influenced chemical reactions in solution [139,177,179,180,181,182] and for the rates of binding between ligands and receptor arrays [139, 141,178,183,184]. In these studies, simulated molecules diffuse in solution; at the moment that a reactant pair, or a ligand and its cognate receptor, come into contact, they undergo a chemical reaction. While diffusing, intermolecular forces are often ignored, although some studies account for these interactions as well [185,186].

To achieve the necessary level of detail, Brownian dynamics simulations usually use very short simulation time steps, often on the order of picoseconds [139]. Adaptive time steps, such that time steps are long when reactants are widely separated and short when they are close, can speed simulations up by several orders of magnitude, but are easy to implement only if there is just one diffusing particle present in the simulation volume [141]. A more sophisticated method that has the same general goal of computational efficiency is called Green's function reaction dynamics [143,187,188] (GFRD). In GFRD, which works with any number of molecules, the system is inspected to see how soon the next molecular collision or reaction could occur. The system is then advanced to that time using a single simulation time step, the event is executed if appropriate, and the cycle repeats. Yet another method, used in a program written by one of us (SSA) called Smoldyn, achieves computational efficiency by modifying the effective radii of simulated molecules so that the same reaction rate is achieved with long simulation time steps as with short ones [45,189,190]. This method does not achieve the same spatial or temporal precision as classical Brownian dynamics or GFRD, but the level of detail is still more than adequate for most biological applications and has been shown to be indistinguishable from more accurate simulations in many cases [191].

Technically, all of these algorithms execute Brownian dynamics. However, the term "Brownian dynamics" is typically used to describe highly detailed studies in which reaction rates, rebinding dynamics, or similar phenomena are found from fundamental molecular properties such as molecular radii and intermolecular forces. In contrast, GFRD and the methods used in Smoldyn are more often used to determine system-level behaviors from known or estimated reaction rates. These are more often called particle-based stochastic simulation methods [192].

MCell is another program that performs particle-based stochastic simulations [144]. Unlike the others, it cannot simulate reactions that occur in free solution, but instead only treats reactions at surfaces. Despite the decrease of versatility, it is still useful for studying a wide variety of biological phenomena [193,194,195]; in particular, it was developed to investigate the neuromuscular junction [196,197,198]. In MCell, surface-bound receptors are not modeled as single molecules as they would be in Smoldyn or GFRD methods, but as a uniform binding probability that applies to an entire surface tile. This decreases the spatial resolution some, but increases the computational efficiency.

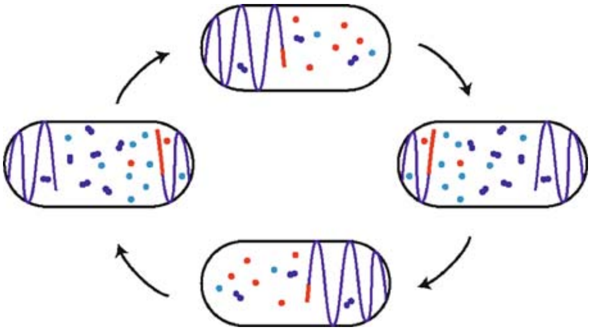
Future Directions

The classic advice of using the right tool for the job is as true in biochemical modeling as it is elsewhere. Several modeling tools have been presented here. Deterministic ordinary differential equation models are simple, easy to use, and can be analyzed with many powerful theoretical and analytical methods. They are the right tool for systems that can be treated as being well-mixed and that are both large enough and sufficiently far from critical points that stochastic effects are unimportant. In contrast, systems that include low copy numbers of important components, and/or that can be triggered by random events, require stochastic modeling methods for their investigation. These include integration of the chemical master equation and random sampling of the stochastic trajectories using the Gillespie algorithm, both of which are exact methods. Approximate methods include τ-leaping stochastic simulations, integration of the chemical Fokker–Planck equation, and sampling with the chemical Langevin equation. Of these, the Gillespie algorithm has proven to be the most popular. Finally, if the system cannot be considered to be well-mixed, then yet different tools are needed. These include spatial variants of the same list of simulation methods, including partial differential equations for deterministic simulations and a spatial Gillespie algorithm for stochastic simulations. Particle-tracking simulation methods allow an even greater degree of detail.

In general, more detailed simulation methods yield more accurate results and are based more closely on underlying processes and less on phenomenological descriptions. However, they are also more computationally intensive and require more model parameters. This parametrization poses a significant problem for current models because the necessary quantitative experimental data are typically only marginally adequate or are completely non-existent. For an ODE model, it is sometimes possible to address this problem by exploring model behaviors over wide ranges of parameter space, from which one can draw phase diagrams that graphically depict how the model behaves for different parameter choices. From this, one can sometimes constrain parameters or gain additional insight into the model; for example, Tyson showed how two enzyme concentrations can be used to regulate the cell cycle, bringing an oocyte from metaphase arrest to autonomous oscillations, and on to growth-controlled cell division [30]. Because of the computational demands of spatial and stochastic models, as well as the richer behavior possibilities, it is much more difficult to explore parameter space with these more complex models. Thus, much work is needed on this topic.

More generally, the mathematical infrastructure for designing and interpreting stochastic models lags far behind that for non-spatial deterministic models. This poses some challenges for theorists. For example, what new theories and graphical tools will help scientists gain intuition into the dynamics of stochastic systems? and what are the controlling elements of stochastic systems? The theory is even more unexplored when spatial organization is considered as well. Nevertheless, spatial considerations are essential because no biological life has been found that is well-mixed; instead, a tremendous amount of biochemical activity involves membranes, polymers, protein scaffolds, large multimeric complexes, and other spatial structures. Theories that address these topics will not be as elegant as those that focus on the chemical master equation, but biology is not always elegant either.

Although research on stochastic modeling of biochemistry grew slowly from the 1950s to the 1990s, the pace has accelerated dramatically during the last 10 to 15 years. This acceleration will likely continue for many more years, in response to the faster computers that become available every year and to the ever-increasing complexity of biochemical data. With this growth, stochastic modeling may open up entire new ways to understand cell biology.



Stochastic Models of Biological Processes, Figure 2

Diagram of the *E. coli* Min system, which is used to position the cell division plane at the cell center. *Dots* represent cytoplasmic proteins, while *curved lines* represent helical membrane-bound protein polymers. *Colors* identify the proteins: *light blue* for MinD bound to ADP, *dark blue* for MinD protein bound to ATP, and *red* for MinE. The system dynamics are summarized in the text of Sect. "Box 1: The E. Coli Min System"

Box 1: The *E. Coli* Min System

The *E. coli* Min system has served as a proving ground for spatial stochastic simulation methods. The Min system is used by *E. coli*, in conjunction with other systems, to position the cell division plane accurately at the cell center [131]. The system is comprised of the proteins MinC, MinD and MinE, which oscillate back and forth across the cell, from one pole to the other, with a period of about 40s (Fig. 2). Of these, only MinD and MinE are required for the oscillation, making this a relatively simple system that exhibits remarkably interesting dynamics. Cytoplasmic MinD proteins bind ATP, dimerize, and polymerize on the inside of the cell membrane to form long helical structures that extend outwards from one of the two cell poles. When MinE binds to the cell-center end of a MinD polymer, it activates ATP hydrolysis which depolymerizes the terminal subunit. As MinE progressively disassembles a MinD polymer at one end of the cell, it reassembles again from the opposite pole to start the next oscillation cycle. The oscillating Min proteins continually inhibit cell division plane formation near the poles using MinC, which colocalizes with MinD, thus only permitting cell division at the cell center.

This system was explored for several years with deterministic reaction-diffusion models [128,199,200]. One of these models, by Howard, Rutenberg, and de Vet [199], was also explored by the same group using a one-dimensional population-based stochastic method [129] (it uses discrete particle numbers and fixed time steps, thus conceptually placing it between the spatial Langevin and spatial Gillespie methods). The authors found that stochastic effects were essential for generating oscillations in some parameter regimes, in a spatial version of stochastic resonance [60,201]. These models helped direct new experiments [202,203,204,205] that clarified the processes of the system.

Building on the prior models and the new experimental data, Huang, Meir, and Wingreen [130] developed a new reaction-diffusion model that was more closely connected with the biology than were previous models and that accounted for several mutant phenotypes. This model became the basis of several stochastic simulations. The spatial Gillespie method was employed by Fange and Elf [132] using their MesoRD program. They showed that a stochastic model can account for a "spotty" phenotype and for oscillations in spherical mutant cells, neither of which can be explained by the deterministic model. The MCell particle-tracking program was used by Kerr and coworkers [133] to show that the Min system alone is insufficient to center the cell division plane with high accuracy.

Yet unexplained with these simulations were convincing experimental results that MinD forms polymers on the cell membrane [205,206,207]. These were explored with another particle-tracking model [208], using a method based on Smoluchowski dynamics [45]. Although this group did simulate spontaneous polymer formation, they observed many randomly oriented short filaments, in contrast to the few helical polymers that are observed experimentally. This inherent difficulty with the reaction-diffusion model [209], whether deterministic or stochastic, has led to several studies that have focused specifically on the polymer dynamics and shapes [134,210,211].

The *E. coli* Min system is already well on its way to becoming the prototypical system for studying spatial biochemical dynamics, much as *E. coli* chemotaxis has become the prototypical system for investigating bacterial signaling.

Acknowledgments

This work was funded by the US Department of Energy.

Bibliography

1. McQuarrie DA (1967) Stochastic approach to chemical kinetics. J Appl Probab 4:413–478
2. Turner TE, Schnell S, Burrage K (2004) Stochastic approaches for modelling in vivo reactions. Comp Biol Chem 28:165–178
3. Raser JM, O'Shea EK (2005) Noise in gene expression: Origins, consequences, and control. Science 309:2010–2013
4. Samoilov MS, Price G, Arkin AP (2006) From fluctuations to phenotypes: The physiology of noise. Sci STKE 2006:re17

5. Rao CV, Wolf DM, Arkin AP (2002) Control, exploitation, and tolerance of intracellular noise. Nature 420:231–237
6. Gillespie DT (2007) Stochastic simulation of chemical kinetics. Ann Rev Phys Chem 58:35–55
7. Wolf DM, Arkin AP (2003) Motifs, modules and games in bacteria. Curr Opin Microbiol 6:125–134
8. Andrews SS, Arkin AP (2006) Simulating cell biology. Curr Biol 16:R523–R527
9. McAdams HH, Arkin A (1999) It's a noisy business! Genetic regulation at the nanomolar scale. Trends Genet 15:65–69
10. Singer RH, Lawrence DS, Ovryn B, Condeelis J (2005) Imaging of gene expression in living cells and tissues. Biomed J Optics 10:051406
11. Levsky JM, Shenoy SM, Pezo RC, Singer RH (2002) Single-cell gene expression profiling. Science 297:836–840
12. Elowitz MB, Levine AJ, Siggia ED, Swain PS (2002) Stochastic gene expression in a single cell. Science 297:1183–1186
13. Raser JM, O'Shea EK (2004) Control of stochasticity in eukaryotic gene expression. Science 304:1811–1814
14. Blake WJ, Kaern M, Cantor CR, Collins JJ (2003) Noise in eukaryotic gene expression. Nature 422:633–637
15. Ozbudak EM, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A (2002) Regulation of noise in the expression of a single gene. Nature Genet 31:69–73
16. Weinberger LS, Burnett JC, Toettcher JE, Arkin AP, Schaffer DV (2005) Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic diversity. Cell 122:169–182
17. Cai L, Friedman N, Xie XS (2006) Stochastic protein expression in individual cells at the single molecule level. Nature 440:358–362
18. Yu J, Xiao J, Ren X, Lao K, Xie XS (2006) Probing gene expression in live cells, one protein molecule at a time. Science 311:1600–1603
19. Golding I, Cox EC (2006) Protein synthesis molecule by molecule. Genome Biol 7:212
20. Fusco D, Accornero N, Lavoie B, Shenoy SM, Blanchard J-M, Singer RH, Bertrand E (2003) Single mRNA molecules demonstrate probabilistic movement in living mammalian cells. Curr Biol 13:161–167
21. Elowitz MB, Leibler S (2000) A synthetic oscillatory network of transcriptional regulators. Nature 403:335–338
22. Sakurai JJ (1994) Modern Quantum Mechanics. Addison-Wesley, Boston
23. Strogatz SH (1994) Nonlinear Dynamics and Chaos. Westview Press, Cambridge
24. Atkins PW (1986) Physical Chemistry. Freeman, New York
25. Tyson JJ, Chen KC, Novak B (2003) Sniffers, buzzers, toggles, and blinkers: dynamics of regulatory and signaling pathways in the cell. Curr Opin Cell Biol 15:221–231
26. Covert MW, Schilling CH, Famili I, Edwards JS, Goryanin II, Selkov E, Palsson BO (2001) Metabolic modeling of microbial strains in silico. Trends Biochem Sci 26:179–186
27. Varma A, Palsson BO (1994) Metabolic flux balancing: Basic concepts, scientific and practical use. Nature Biotech 12:994–998
28. Kauffman KJ, Prakash P, Edwards JS (2003) Advances in flux balance analysis. Curr Opin Biotech 14:491–496
29. Fell D (1997) Understanding the Control of Metabolism. Portland Press, London
30. Tyson JJ (1991) Modeling the cell division cycle: cdc2 and cyclin interactions. Proc Natl Acad Sci USA 88:7328–7332
31. Chen KC, Calzone L, Csikasz-Nagy A, Cross FR, Novak B, Tyson JJ (2004) Integrative analysis of cell cycle control in budding yeast. Mol Biol Cell 15:3841–3862
32. Goldbeter A (2002) Computational approaches to cellular rhythms. Nature 420:238–245
33. van Zon JS, Lubensky DK, Altena PRH, ten Wolde PR (2007) An allosteric model of circadian KaiC phosphorylation. Proc Natl Acad Sci USA 104:7420–7425
34. Reinitz J, Mjolsness E, Sharp DH (1995) Model for cooperative control of positional information in Drosophila by bicoid and maternal hunchback. Exp J Zool 271:47–56
35. von Dassow G, Meir E, Munro EM, Odell GM (2000) The segment polarity network is a robust developmental module. Nature 406:188–192
36. Kellershohn N, Laurent M (2001) Prion diseases: dynamics of the infection and properties of the bistable transition. Biophys J 81:2517–2529
37. Ferrell JEJ, Machleder EM (1998) The biochemical basis of an all-or-none cell fate switch in Xenopus oocytes. Science 280:895–898
38. Huang C-YF, Ferrell JEJ (1996) Ultrasensitivity in the mitogen-activated protein kinase cascade. Proc Natl Acad Sci USA 93:10078–10083
39. Laurent M, Kellershohn N (1999) Multistability: a major means of differentiation and evolution in biochemical systems. Trends Biochem Sci 24:418–422
40. Samoilov MS, Arkin AP (2006) Deviant effects in molecular reaction pathways. Nature Biotech 24:1235–1240
41. van Kampen NG (1992) Stochastic Processes in Physics and Chemistry. Elsevier, Amsterdam
42. Haseltine EL, Rawlings JB (2005) On the origins of approximations for stochastic chemical kinetics. Chem J Phys 123:164115
43. Gillespie DT (1992) A rigorous derivation of the chemical master equation. Physica A 188:404–425
44. Rohwer JM, Postma PW, Kholodenko BN, Westerhoff HV (1998) Implications of macromolecular crowding for signal transduction and metabolite channeling. Proc Natl Acad Sci USA 95:10547–10552
45. Andrews SS, Bray D (2004) Stochastic simulation of chemical reactions with spatial resolution and single molecule detail. Phys Biol 1:137–151
46. Munsky B, Khammash M (2006) The finite state projection algorithm for the solution of the chemical master equation. Chem J Phys 124:044104
47. Peles S, Munsky B, Khammash M (2006) Reduction and solution of the chemical master equation using time scale separation and finite state projection. Chem J Phys 125:204104
48. Kuwahara H, Myers CJ, Samoilov MS, Barker NA, Arkin AP (2006) Automated abstraction methodology for genetic regulatory networks. Trans Comput Syst Biol 6:150–175
49. Hegland M, Burden C, Santoso L, MacNamara S, Booth H (2007) A solver for the stochastic master equation applied to gene regulatory networks. Comp J Appl Math 205:708–724
50. Nedea SV, Jansen APJ, Lukkien JJ, Hilbers PAJ (2003) Infinitely fast diffusion in single-file systems. Phys Rev E 67:046707
51. Chatterjee A, Vlachos DG (2006) Multiscale spatial Monte Carlo simulations: Multigriding, computational singular per-

turbation, and hierarchical stochastic closures. Chem J Phys 124:064110

52. Ambjörnsson T, Banik SK, Lomholt MA, Metzler R (2007) Master equation approach to DNA breathing in heteropolymer DNA. Phys Rev E 75:021908
53. Altan-Bonnet G, Libchaber A, Krichevsky O (2003) Bubble dynamics in double-stranded DNA. Phys Rev Lett 90:138101
54. Lattanzi G, Maritan A (2001) Master equation approach to molecular motors. Phys Rev E 64:061905
55. Wang H-Y, Elston T, Mogilner A, Oster G (1998) Force generation in RNA polymerase. Biophys J 74:1186–1202
56. Peskin CS, Odell GM, Oster GF (1993) Cellular motions and thermal fluctuations: the Brownian ratchet. Biophys J 65:316–324
57. Paulsson J, Ehrenberg M (2000) Random signal fluctuations can reduce random fluctuations in regulated component of chemical regulatory networks. Phys Rev Lett 84:5447–5450
58. Paulsson J, Berg OG, Ehrenberg M (2000) Stochastic focusing: fluctuation-enhanced sensitivity of intracellular regulation. Proc Natl Acad Sci USA 97:7148–7153
59. Berg OG, Paulsson J, Ehrenberg M (2000) Fluctuations in repressor control: thermodynamic constraints on stochastic focusing. Biophys J 79:2944–2953
60. Li H, Hou Z, Xin H (2005) Internal noise stochastic resonance for intracellular calcium oscillations in a cell system. Phys Rev E 71:061916
61. Samoilov M, Plyasunov S, Arkin AP (2005) Stochastic amplification and signaling in enzymatic futile cycles through noise-induced bistability with oscillations. Proc Natl Acad Sci USA 102:2310–2315
62. Gillespie DT (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. Comp J Phys 22:435–450
63. Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. Phys J Chem 81:2340–2361
64. Gibson MA, Bruck J (2000) Efficient exact stochastic simulation of chemical systems with many species and many channels. Phys J Chem A 104:1876–1889
65. Lok L, Brent R (2005) Automatic generation of cellular reaction networks with Molecularizer 1.0. Nature Biotech 23:131–136
66. Cao Y, Li H, Petzold L (2004) Efficient formulation of the stochastic simulation algorithm for chemically reacting systems. Chem J Phys 121:4059–4067
67. McCollum JM, Peterson GD, Cox CD, Simpson ML, Samatova NF (2006) The sorting direct method for stochastic simulation of biochemical systems with varying reaction execution behavior. Comp Biol Chem 30:39–49
68. Plyasunov S, Arkin AP (2007) Efficient stochastic sensitivity analysis of discrete event systems. Comput J Phys 221:724–738
69. Bardwell L (2004) A walk-through of the yeast mating pheromone response pathway. Peptides 25:1465–1476
70. Morton-Firth CJ, Bray D (1998) Predicting temporal fluctuations in an intracellular signalling pathway. Theor J Biol 192:117–128
71. LeNovère N, Shimizu TS (2001) StochSim: modelling of stochastic biomolecular processes. Bioinformatics 17:575–576
72. Lu T, Volfson D, Tsimring L, Hasty J (2004) Cellular growth and division in the Gillespie algorithm. Syst Biol 1:121–128
73. McAdams H, Arkin A (1998) Simulation of prokaryotic genetic circuits. Annu Rev Biophys Biomol Struct 27:199–224
74. Paulsson J (2004) Summing up the noise in gene networks. Nature 427:415–418
75. McAdams HH, Arkin A (1997) Stochastic mechanisms in gene expression. Proc Natl Acad Sci USA 94:814–819
76. Arkin A, Ross J, McAdams HH (1998) Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. Genetics 149:1633–1648
77. Wolf DM, Vazirani VV, Arkin AP (2005) Diversity in times of adversity: probabilistic strategies in microbial survival games. Theor J Biol 234:227–253
78. Fiering S, Whitelaw E, Martin DIK (2000) To be or not to be active: the stochastic nature of enhancer action. BioEssays 22:381–387
79. Barkai N, Leibler S (2000) Biological rhythms: Circadian clocks limited by noise. Nature 403:267–268
80. Thattai M, van Oudenaarden A (2004) Stochastic gene expression in fluctuating environments. Genetics 167:523–530
81. Kussell E, Leibler S (2005) Phenotypic diversity, population growth, and information in fluctuating environments. Science 309:2075–2078
82. Pedraza JM, van Oudenaarden A (2005) Noise propagation in gene networks. Science 307:1965–1969
83. Swain PS, Elowitz MB, Siggia ED (2002) Intrinsic and extrinsic contributions to stochasticity in gene expression. Proc Natl Acad Sci USA 99:12795–12800
84. Mettetal JT, Muzzey D, Pedraza JM, Ozbudak EM, van Oudenaarden A (2006) Predicting stochastic gene expression dynamics in single cells. Proc Natl Acad Sci USA 103:7304–7309
85. Kierzek AM, Zaim J, Zielenkiewicz P (2001) The effect of transcription and translation initiation frequencies on the stochastic fluctuations in prokaryotic gene expression. Biol J Chem 276:8165–8172
86. Peccoud J, Ycart B (1995) Markovian modeling of gene-product synthesis. Theor Popul Biol 48:222–234
87. Rosenfeld N, Young JW, Alon U, Swain PS, Elowittz MB (2005) Gene regulation at the single-cell level. Science 307:1962–1965
88. Kitano H (2004) Biological robustness. Nature Rev Genet 5:826–837
89. Alon U, Surette MG, Barkai N, Leibler S (1999) Robustness in bacterial chemotaxis. Nature 397:168–171
90. Barkai N, Leibler S (1997) Robustness in simple biochemical networks. Nature 387:913–917
91. Stelling J, Sauer U, Szallasi Z, Doyle FJI, Doyle J (2004) Robustness of cellular functions. Cell 118:675–685
92. Vilar JMG, Kueh HY, Barkai N, Leibler S (2002) Mechanisms of noise-resistance in genetic oscillators. Proc Natl Acad Sci USA 99:5988–5992
93. Aldana M, Cluzel P (2003) A natural class of robust networks. Proc Natl Acad Sci USA 100:8710–8714
94. Thattai M, van Oudenaarden A (2002) Attenuation of noise in ultrasensitive signaling cascades. Biophys J 82:2943–2950
95. Yi T-M, Huang Y, Simon MI, Doyle J (2000) Robust perfect adaptation in bacterial chemotaxis through integral feedback control. Proc Natl Acad Sci USA 97:4649–4653
96. Fraser HB, Hirsh AE, Giaever G, Kumm J, Eisen MB (2004) Noise minimization in eukaryotic gene expression. PLoBiol S 2:1–5
97. Voigt CA, Wolf DM, Arkin AP (2005) The Bacillus subtilis sin operon: an evolvable network motif. Genetics 169:1187–1202

98. Cao Y, Gillespie DT, Petzold LR (2006) Efficient step size selection for the tau-leaping simulation method. Chem J Phys 124:044109
99. Gillespie DT, Petzold LR (2003) Improved leap-size selection for accelerated stochastic simulation. Chem J Phys 119:8229–8234
100. Gillespie DT (2001) Approximate accelerated stochastic simulation of chemically reacting systems. Chem J Phys 115:1716–1733
101. Cao Y, Gillespie DT, Petzold LR (2005) Avoiding negative populations in explicit Poisson tau-leaping. Chem J Phys 123:054104
102. Chatterjee A, Mayawala K, Edwards JS, Vlachos DG (2005) Time accelerated Monte Carlo simulations of biological networks using the binomial tau-leap method. Bioinformatics 21:2136–2137
103. Pettigrew MF, Resat H (2007) Multinomial tau-leaping method for stochastic kinetic simulations. Chem J Phys 126:084101
104. Zwanzig R (2001) A chemical Langevin equation with non-Gaussian noise. Phys J Chem B 105:6472–6473
105. Gillespie DT (1996) The multivariate Langevin and Fokker–Planck equations. Am Phys J 64:1246–1257
106. Gillespie DT (2000) The chemical Langevin equation. Chem J Phys 113:297–306
107. Gillespie DT (2002) The chemical Langevin and Fokker–Planck equations for the reversible isomerization reaction. Phys J Chem A 106:5063–5071
108. Wang H, Peskin CS, Elston TC (2003) A robust numerical algorithm for studying biomolecular transport processes. Theor J Biol 221:491–511
109. Xing J, Wang H, Oster G (2005) From continuum Fokker–Planck models to discrete kinetic models. Biophys J 89:1551–1563
110. Tao Y (2004) Intrinsic noise, gene regulation and steady-state statistics in a two-gene network. Theor J Biol 231:563–568
111. van der Mee CVM, Zweifel PF (1987) A Fokker–Planck equation for growing cell populations. Math J Biol 25:61–72
112. Sato K, Kaneko K (2006) On the distribution of state values of reproducing cells. Phys Biol 3:74–82
113. Hill NA, Häder D-P (1997) A biased random walk model for the trajectories of swimming micro-organisms. Theor J Biol 186:503–526
114. Schienbein M, Gruler H (1993) Langevin equation, Fokker–Planck equation and cell migration. Bull Math Biol 55:585–608
115. Xing J, Liao J-C, Oster G (2005) Making ATP. Proc Natl Acad Sci USA 102:16539–16546
116. Elston TC, Oster G (1997) Protein turbines I: the bacterial flagellar motor. Biophys J 73:703–721
117. Allen RJ, Frenkel D, ten Wolde PR (2006) Simulating rare events in equilibrium or nonequilibrium stochastic systems. Chem J Phys 124:024102
118. Rathinam M, Petzold LR, Cao Y, Gillespie DT (2003) Stiffness in stochastic chemically reacting systems: The implicit tau-leaping method. Chem J Phys 119:12784
119. Cao Y, Gillespie DT, Petzold LR (2005) The slow-scale stochastic simulation algorithm. Chem J Phys 122:014116
120. Cao Y, Gillespie DT, Petzold LR (2005) Accelerated stochastic simulation of the stiff enzyme-substrate reaction. Chem J Phys 123:144917
121. Rao CV, Arkin AP (2003) Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the Gillespie algorithm. Chem J Phys 118:4999–5010
122. Adalsteinsson D, McMillen D, Elston TC (2004) Biochemical network stochastic simulator (BioNetS): software for stochastic modeling of biochemical networks. Bioinformatics BMC 5:24
123. Vasudeva K, Bhalla US (2004) Adaptive stochastic-deterministic chemical kinetic simulations. Bioinformatics 20:78–84
124. Baumeister W (2002) Electron tomography: towards visualizing the molecular organization of the cytoplasm. Curr Opin Struct Biol 12:679–684
125. Gierer A, Meinhardt H (1972) A theory of biological pattern formation. Biol Cyber 12:30–39
126. Maini PK, Painter KJ, Chau HNP (1997) Spatial pattern formation in chemical and biological systems. Chem J Soc Faraday Trans 93:3601–3610
127. Gurdon JB, Bourillot P-Y (2001) Morphogen gradient interpretation. Nature 413:797–803
128. Meinhardt H, de Boer PAJ (2001) Pattern formation in *Escherichia coli*: A model for the pole-to-pole oscillations of Min proteins and the localization of the division site. Proc Natl Acad Sci USA 98:14202–14207
129. Howard M, Rutenberg AD (2003) Pattern formation inside bacteria: fluctuations due to the low copy number of proteins. Phys Rev Lett 90:128102
130. Huang KC, Meir Y, Wingreen NS (2003) Dynamic structures in *Escherichia coli*: spontaneous formation of MinE rings and MinD polar zones. Proc Natl Acad Sci USA 100:12724–12728
131. Lutkenhaus J (2007) Assembly and dynamics of the bacterial MinCDE system and spatial regulation of the Z ring. Ann Rev Biochem 76:14.11–14.24
132. Fange D, Elf J (2006) Noise-induced Min phenotypes in *E coli*. PLoComp S Biol 2:637–648
133. Kerr RA, Levine H, Sejnowski TJ, Rappel W-J (2006) Division accuracy in a stochastic model of Min oscillations in *Escherichia coli*. Proc Natl Acad Sci USA 103:347–352
134. Cytrynbaum E, Marshall BDL (2007) A multi-stranded polymer model explains MinDE dynamics in *E coli* cell division. Biophys J 93:1134–1150
135. Bray D (1998) Signaling complexes: biophysical constraints on intracellular communication. Annu Rev Biophys Biomol Struct 27:59–75
136. Slepchenko BM, Schaff JC, Carson JH, Loew LM (2002) Computational cell biology: Spatiotemporal simulation of cellular events. Annu Rev Biophys Biomol Struct 31:423–441
137. Meyers J, Craig J, Odde DJ (2006) Potential for control of signaling pathways via cell size and shape. Curr Biol 16:1685–1693
138. Rao CV, Kirby JR, Arkin AP (2005) Phosphatase localization in bacterial chemotaxis: divergent mechanisms, convergent principles. Phys Biol 2:148–158
139. Agmon N, Edelstein AL (1997) Collective binding properties of receptor arrays. Biophys J 72:1582–1594
140. Lagerholm BC, Thompson NL (1998) Theory for ligand rebinding at cell membrane surfaces. Biophys J 74:1215–1228
141. Andrews SS (2005) Serial rebinding of ligands to clustered receptors as exemplified by bacterial chemotaxis. Phys Biol 2:111–122

142. Elf J, Ehrenberg M (2004) Spontaneous separation of bi-stable biochemical systems into spatial domains of opposite phases. Syst Biol 1:230–236
143. van Zon JS, ten Wolde PR (2005) Green's function reaction dynamics: A particle-based approach for simulating biochemical networks in time and space. Chem J Phys 123:234910
144. Stiles JR, Bartol TM (2001) Monte Carlo methods for simulating realistic synaptic microphysiology using MCell. In: De Schutter E (ed) Computational Neuroscience: Realistic Modeling for Experimentalists. Press CRC, Boca Raton
145. Dab D, Boon J-P, Li Y-X (1991) Lattice-gas automata for coupled reaction-diffusion equation. Phys Rev Lett 66:2535–2539
146. Ermentrout GB, Edelstein-Keshet L (1993) Cellular automata approaches to biological modeling. Theor J Biol 160:97–133
147. Duke TAJ, LeNovère N, Bray D (2001) Conformational spread in a ring of proteins: a stochastic approach to allostery. Mol J Biol 308:541–553
148. Goldman J, Andrews SS, Bray D (2004) Size and composition of membrane protein clusters predicted by Monte Carlo analysis. Eur Biophys J 33:506–512
149. Grima R, Schnell S (2006) A systematic investigation of the rate laws valid in intracellular environments. Biophys Chem 124:1–10
150. Turing AM (1990) The chemical basis of morphogenesis. Bull Math Biol 52:153–197
151. Cross MC, Hohenberg PC (1993) Pattern formation outside of equilibrium. Rev Mod Phys 65:851–1123
152. Slepchenko B, Schaff J, Macara I, Loew LM (2003) Quantitative cell biology with the Virtual Cell. Cell TRENDS Biol 13:570–576
153. Fink CC, Slepchenko B, Moraru II, Watras J, Schaff JC, Loew LM (2000) An image-based model of calcium waves in differentiated neuroblastoma cells. Biophys J 79:163–183
154. Fink CC, Slepchenko B, Moraru II, Schaff J, Watras J, Loew LM (1999) Morphological control of inositol-1,4,5-triphosphate-dependent signals. Cell J Biol 147:929–935
155. Hernjak N, Slepchenko B, Fernald K, Fink CC, Fortin D, Moraru II, Watras J, Loew LM (2005) Modeling and analysis of calcium signaling events leading to long-term depression in cerebellar Purkinje cells. Biophys J 89:3790–3806
156. Baras F, Malek-Mansour M (1996) Reaction-diffusion master equation: A comparison with microscopic simulations. Phys Rev E 54:6139–6148
157. Stundzia AB, Lumsden CJ (1996) Stochastic simulation of coupled reaction-diffusion processes. Comput J Phys 127:196–207
158. Nicolis G, Malek-Mansour M (1980) Systematic analysis of the multivariate master equation for a reaction-diffusion system. Stat J Phys 22:495–512
159. Kruse K, Elf J (2006) Kinetics in spatially extended systems. In: Szallasi Z, Stelling J, Periwal V (eds) System Modeling in Cell Biology From Concepts to Nuts and Bolts. Press MIT, Cambridge, pp 177–198
160. Hynes JT (1985) The theory of reactions in solution. In: Baer M (ed) Theory of Chemical Reaction Dynamics. Press CRC, Boca Raton, pp 171–234
161. Cohen B, Huppert D, Agmon N (2000) Non-exponential Smoluchowski dynamics in fast acid-base reaction. Am J Chem Soc 122:9838–9839
162. Noyes RM (1955) Kinetics of competitive processes when reactive fragments are produced in pairs. Am J Chem Soc 77:2042–2045
163. Pines E, Huppert D (1988) Geminate recombination in excited-state proton transfer reactions: Numerical solution of the Debye–Smoluchowski equation with backreaction and comparison with experimental results. Chem J Phys 88:5620–5630
164. Verkman AS (2002) Solute and macromolecule diffusion in cellular aqueous compartments. Trends Biochem Sci 27:27–33
165. Schnell S, Turner TE (2004) Reaction kinetics in intracellular environments with macromolecular crowding: simulations and rate laws. Prog Biophys Mol Biol 85:235–260
166. Fulton AB (1982) How crowded is the cytoplasm? Cell 30:345–347
167. Bernstein D (2005) Simulating mesoscopic reaction-diffusion systems using the Gillespie algorithm. Phys Rev E 71:041103
168. Ander M, Beltrao P, Di Ventura B, Ferkinghoff-Borg J, Foglierini M, Kaplan A, Lemerle C, Tomás-Oliveira I, Serrano L (2004) SmartCell, a framework to simulate cellular processes that combines stochastic approximation with diffusion and localisation: analysis of simple networks. Syst Biol 1: 129–138
169. Fricke T, Schnakenberg J (1990) Monte-Carlo simulation of an inhomogeneous reaction-diffusion system in the biophysics of receptor cells. Z Phys B 83:277–284
170. Isaacson SA, Peskin CS (2006) Incorporating diffusion in complex geometries into stochastic chemical kinetics simulations. Sci SIAMJ Comput 28:47–74
171. Hattne J, Fange D, Elf J (2005) Stochastic reaction-diffusion simulation with MesoRD. Bioinformatics 21:2923–2924
172. Frenkel D, Smit B (2002) Understanding molecular simulation: from algorithms to applications. Academic, San Diego
173. Berg HC (1993) Random Walks in Biology. Princeton Univ Press, Princeton
174. Gillespie DT (1996) The mathematics of Brownian motion and Johnson noise. Am Phys J 64:225–240
175. Rice SA (1985) Diffusion Limited Reactions. Elsevier, Amsterdam
176. Ermak DL, McCammon JA (1978) Brownian dynamics with hydrodynamic interactions. Chem J Phys 69:1352–1360
177. Northrup SH, Allison SA, McCammon JA (1984) Brownian dynamics simulation of diffusion-influenced bimolecular reactions. Chem J Phys 80:1517–1524
178. Northrup SH (1988) Diffusion-controlled ligand binding to multiple competing cell-bound receptors. Phys J Chem 92:5847–5850
179. Northrup SH, Erickson HP (1992) Kinetics of protein-protein association explained by Brownian dynamics computer simulation. Proc Natl Acad Sci USA 89:3338–3342
180. Edelstein AL, Agmon N (1993) Brownian dynamics simulations of reversible reactions in one dimension. Chem J Phys 99:5396–5404
181. Oh C, Kim H, Shin KJ (2002) Excited-state diffusion-influenced reversible association-dissociation reaction: Brownian dynamics simulation in three dimensions. Chem J Phys 117:3269–3277
182. Kim H, Yang M, Shin KJ (1999) Dynamic correlation effect in reversible diffusion-influenced reactions: Brownian dynamics simulation in three dimensions. Chem J Phys 111:1068–1075
183. Agmon N, Edelstein A (1995) Geometric and many-particle aspects of transmitter binding. Biophys J 68:815–825

184. Edelstein AL, Agmon N (1997) Brownian simulation of many-particle binding to a reversible receptor array. Comput J Phys 132:260–275
185. Agmon N, Szabo A (1990) Theory of reversible diffusion-influenced reactions. Chem J Phys 92:5270–5284
186. Kim H, Shin KJ (1999) Exact solution of the reversible diffusion-influenced reaction for an isolated pair in three dimensions. Phys Rev Lett 82:1578–1581
187. van Zon JS, ten Wolde PR (2005) Simulating biochemical networks at the particle level in time and space: Green's function reaction dynamics. Phys Rev Lett 94:128103
188. van Zon JS, Morelli MJ, Tanase-Nicola S, ten Wolde PR (2006) Diffusion of transcription factors can drastically enhance the noise in gene expression. Biophys J 91:4350–4367
189. Lipkow K (2006) Changing cellular location of CheZ predicted by molecular simulations. Comp PLOS Biol 2:301–310
190. Lipkow K, Andrews SS, Bray D (2004) Simulated diffusion of CheYp through the cytoplasm of *E coli*. J Bact 187:45–53
191. Tournier AL, Fitzjohn PW, Bates PA (2006) Probability-based model of protein-protein interactions on biological timescales. Algorithms Molec Biol 1:25
192. Tolle DP, Le Novère N (2006) Particle-based stochastic simulation in systems biology. Curr Bioinformatics 1:1–6
193. Franks KM, Bartol TM, Sejnowski TJ (2002) A Monte Carlo model reveals independent signaling at central glutamatergic synapses. Biophys J 83:2333–2348
194. Coggan JS, Bartol TM, Esquenazi E, Stiles JR, Lamont S, Martone ME, Berg DK, Ellisman MH, Sejnowski TJ (2005) Evidence for ectopic neurotransmission at a neuronal synapse. Science 309:446–451
195. Koh X, Srinivasan B, Ching HS, Levchenko A (2006) A 3D Monte Carlo analysis of the role of dyadic space geometry in spark generation. Biophys J 90:1999–2014
196. Stiles JR, van Helden D, Thomas J, Bartol M, Salpeter EE, Salpeter MM (1996) Miniature endplate current rise times < 100 microseconds from improved dual recordings can be modeled with passive acetylcholine diffusion from a synaptic vesicle. Proc Natl Acad Sci USA 93:5747–5752
197. Stiles JR, Kovyazina IV, Salpeter EE, Salpeter MM (1999) The temperature sensitivity of miniature endplate currents is mostly governed by channel gating: Evidence from optimized recordings and Monte Carlo simulations. Biophys J 77:1177–1187
198. Bartol TMJ, Land BR, Salpeter EE, Salpeter MM (1991) Monte Carlo simulation of miniature endplate current generation in the vertebrate neuromuscular junction. Biophys J 59:1290–1307
199. Howard M, Rutenberg AD, de Vet S (2001) Dynamic compartmentalization of bacteria: accurate division in *E coli*. Phys Rev Lett 87:278102
200. Kruse K (2002) A dynamic model for determining the middle of *Escherichia coli*. Biophys J 82:618–627
201. Wio HS (1996) Stochastic resonance in a spatially extended system. Phys Rev E 54:R3075–R3078
202. Hu Z, Gogol EP, Lutkenhaus J (2002) Dynamic assembly of MinD on phospholipid vesicles regulated by ATP and MinE. Proc Natl Acad Sci USA 99:6761–6766
203. Hu Z, Saez C, Lutkenhaus J (2003) Recruitment of MinC, an inhibitor of Z-ring formation, to the membrane in *Escherichia coli*: role of MinD and MinE. J Bact 185:196–203
204. Shih Y-L, Fu X, King GF, Le T, Rothfield L (2002) Division site placement in *E coli*: mutations that prevent formation of the MinE ring lead to loss of the normal midcell arrest of growth of polar MinD membrane domains. EMBOJ 21:3347–3357
205. Shih Y-L, Le T, Rothfield L (2003) Division site selection in *Escherichia coli* involves dynamic redistribution of Min proteins within coiled structures that extend between the two cell poles. Proc Natl Acad Sci USA 100:7865–7870
206. Shih Y-L, Kawagishi I, Rothfield L (2005) The MreB and Min cytoskeletal-like systems play independent roles in prokaryotic polar differentiation. Mol Microbiol 58:917–928
207. Suefuji K, Valluzzi R, RayChaudhuri D (2002) Dynamic assembly of MinD into filament bundles modulated by ATP, phospholipids, and MinE. Proc Natl Acad Sci USA 99:16776–16781
208. Pavin N, Paljetak C, Krstic V (2006) Min-protein oscillations in *Escherichia coli* with spontaneous formation of two-stranded filaments in a three-dimensional stochastic reaction-diffusion model. Phys Rev E 73:021904
209. Adelman JL, Andrews SS (2004) Intracellular pattern formation: A spatial stochastic model of bacterial division site selection proteins MinProc CDE. Complex Systems Summer School Final Project Papers, Santa Fe Institute, Santa Fe
210. Drew DA, Osborn MJ, Rothfield LI (2005) A polymerization-depolymerization model that accurately generates the self-sustained oscillatory system involved in bacterial division site placement. Proc Natl Acad Sci USA 102:6114–6118
211. Andrews SS, Arkin AP (2007) A mechanical explanation for cytoskeletal rings and helices in bacteria. Biophys J 93:1872–1884
212. Lippincott-Schwartz J, Snapp E, Kenworthy A (2001) Studying protein dynamics in living cells. Nat Rev Mol Cell Biol 2:444–456

Stochastic Noises, Observation, Identification and Realization with

GIORGIO PICCI
Department of Information Engineering,
University of Padua, Padua, Italy

Article Outline

Glossary

Wiener process A *Wiener process* is a continuous time stochastic process $w = \{w(t)\,;\, t \in \mathbb{R}\}$ with continuous sample paths and stationary independent increments of zero mean and finite variance. The variance of the increments $w(t) - w(s)$ must then have the form $\sigma^2(t-s)$ and $w(t) - w(s)$ must have a Gaussian distribution with mean zero (Levy's theorem [30]). The Wiener process is *normalized* (or *standard*) if $\sigma^2 = 1$.

A *p-dimensional* Wiener process is a vector stochastic process having p components, w_k; $k = 1, \dots, p$, which are independent Wiener processes. A *wide-sense* p-dimensional normalized Wiener process is a continuous time stochastic process $w = \{w(t)\,;\ t \in \mathbb{R}\}$ which is continuous in mean square and has stationary orthogonal increments; i. e.,

$$\mathbb{E}\{[w(t_1) - w(s_1)][w(t_2) - w(s_2)]'\} = I_p|[s_1,\ t_1] \cap [s_2,\ t_2]|$$

where $\mathbb{E}$ denotes expectation, I_p is the $p \times p$ identity matrix, the prime denotes transpose and $|\cdot|$ denotes length (Lebesgue measure). This is a weaker concept than that of a Wiener process.

A Wiener process with $w(0) = 0$ is sometimes called a *Brownian motion process*. Usually a Brownian motion is only defined on the half line $\mathbb{R}_+$.

Wiener processes and stochastic integrals with respect to a Wiener process, introduced by Wiener and Itô, are the basic building blocks of *stochastic calculus* [14,25]. Since stochastic integrals only depend on the increments of w, it is immaterial whether an arbitrary random vector is added to w. The equivalence class of (wide sense) Wiener processes differing from each other by the addition of an arbitrary (constant) random vector of finite variance (e. g. $w(0)$), will still be called a Wiener process and denoted by the symbol dw.

White noise Continuous time *white noise* is the formal time derivative of a Wiener process. This derivative must be understood in a suitable distributional sense since it fails to exist as a limit in any of the standard senses of probability theory.

Much more standard is the notion of *discrete-time* (strict or wide-sense) normalized *white noise process*. It is a sequence of independent equally distributed (respectively orthogonal) random vectors $w = \{w(t)\,;\ t \in \mathbb{Z}\}$ with unit variance; i. e.,

$$\mathbb{E}\{w(t)w(s)'\} = I\delta_{ts} := \begin{cases} I & \text{if } s = t \\ 0 & \text{if } s \neq t\,. \end{cases}$$

This last condition alone characterizes wide-sense white noise. We shall always require that w should have zero mean.

Stochastic differential and difference equations

Stochastic differential equations are equations that define pathwise a continuous time stochastic process x by "local" evolution laws of the type

$$\mathrm{d}x(t) = f(x)\mathrm{d}t + G(x)\mathrm{d}w(t)\,,$$

where dw is a vector Wiener process. The equation should in reality be interpreted as an integral equation, the last term being a stochastic integral in the sense of Itô, see [14,25]. Under certain growth conditions on the coefficients the equation can be shown to have a unique solution which, in case f and G depend pointwise on $x(t)$, is a continuous Markov process, in fact a *diffusion process*.

In discrete time ($t \in \mathbb{Z}$) a *stochastic difference equation* is an object of the type

$$x(t+1) = f(x(t))\mathrm{d}t + G(x(t))w(t)$$

where now w is white noise. The solution of an equation of this type is a (discrete-time) Markov process.

Stochastic dynamical systems In continuous time, a (finite-dimensional) *stochastic system* is a pair (x, y) of vector, say n and m dimensional, stochastic processes satisfying equations of the type

$$\begin{aligned} \mathrm{d}x(t) &= f(x(t))\mathrm{d}t + G(x(t))\mathrm{d}w(t) \\ \mathrm{d}y(t) &= h(x(t))\mathrm{d}t + J(x(t))\mathrm{d}w(t)\,. \end{aligned}$$

The process x, which is Markov, is called the *state* of the system while y is the *output* process. The differential notation in the second equation is merely to allow a possible additive white noise component in the output process so y is actually the integral of the variable which would physically be called the output of the system. If there is no additive white noise component ($J \equiv 0$) this trick is unnecessary and the output variable is simply expressed as a memoryless function of the state, $\mathrm{d}y/\mathrm{d}t = h(x(t))$. A similar definition, with the obvious modifications, serves to introduce the concept of discrete-time stochastic systems.

The probabilistic essence of the concept of stochastic system is more generally captured in terms of *conditional independence* of the sigma-algebras induced by the underlying processes. From this it can be clearly seen that the state process at time t plays the role of *dynamic memory* of the system: the future and the past of the output and of the state sigma-algebras are conditionally independent given the present state $x(t)$. For a formal definition the reader is for example referred to p. 219 in [51]. We shall introduce these concepts in a wide-sense context later on.

Introduction

In this article we discuss some general ideas which motivate the use of stochastic dynamical models in applied sciences. We discuss the inherent mathematical problem of stochastic model building, called *stochastic realization* and the statistical problem of *(dynamic) system identification*, i. e. procedures for estimating dynamic stochastic models starting from observed data. We shall in particular discuss linear stochastic state-space models with random inputs, their construction and the relevant identification techniques. We shall assume that the reader has some general background on stationary stochastic processes, statistics [12] and linear system theory as for example exposed in the textbook [24].

There are no physical laws telling us how to obtain stochastic descriptions of nature. The laws of mechanics, electromagnetism, fluid dynamics etc. are, by nature, essentially deterministic. A fundamental empirical observation is that stochastic models generally come about as *aggregate* descriptions of complicated large deterministic systems. Think for example of describing mathematically the outcome of a dice-throwing experiment. Doing this by the laws of physics involves an enormous number of complicated factors such as the mechanics of a cubic-shaped rigid body flying in the air after an initial impulse applied on a specific region of one of its six faces. To describe the trajectory (i. e. the motion of the barycenter, the body orientation and the angular momenta) one should take into account, besides gravity, lift phenomena due to portance, drag, added-mass effects etc., describing the interaction with the surrounding fluid. Then one should describe the discontinuous mechanics of the dice landing on an elastic surface etc. Modeling all of this by the known laws of physics is perhaps possible and we may in principle be able to set up a set of algebraic-differential equations allowing us to predict exactly (or "almost exactly") which upper face of the dice will eventually show. This however, provided we knew an enormous number of geometrical and physical parameters of the system such as, for example, the dimensions of the dice, its initial position and orientation, its mass and moments of inertia, the location, direction and strength of the initial impulse, the density of the air surrounding the trajectory (depending on the temperature, humidity etc.), the possible air convection phenomena, the mechanical structure and geometry of the landing medium, etc. Indeed, this predictive model based on the (deterministic) laws of physics would require a very large number of equations and involve an unimaginable detailed prior knowledge of the parameters of the experiment. Trivially, the rudimentary stochastic model consisting of a probability distribution on the six possible outcomes $\{1, 2, \dots, 6\}$, although not allowing an "exact", i. e. "deterministic" prediction of the outcome, is incomparably simpler. One could venture to say that most stochastic models used in science and engineering come about, for the same reason, as simple aggregate descriptions of complicated deterministic systems. Thermodynamics, to name just one of the most conspicuous instances of this phenomenon, can be seen as stochastic aggregation of large ensembles of microscopic particles evolving in time according to the deterministic laws of mechanics [28,43].

Simplification is payed in terms of uncertainty and the predictions based on stochastic aggregate models are by nature uncertain. The modeling process hence requires us to quantify the uncertainty of aggregate modes. In a more precise sense, we should investigate their mathematical structure and describe how to construct them. This will be one of the main concerns of this article.

A Brownian Particle in a Heat Bath

The process of constructing aggregate models can be elucidated by a famous example, the *Ford–Kac–Mazur model* [20] which is probably the simplest explicitly solvable example of a problem of this nature.

Consider the problem of describing the dynamics of a particle, called the *Brownian particle*, coupled to a "heat bath" consisting of a very large number of identical particles obeying the laws of classical Hamiltonian mechanics and moving under the influence of a quadratic potential. Mathematically, the ensemble can be described as a large number of coupled harmonic oscillators. The Brownian particle is the only particle of the ensemble which is assumed to be accessible to external observation. One assumes that the macroscopic observables of the system are the relative position $q(t)$ and/or the momentum $p(t)$ of the Brownian particle, at each instant of time.

It is shown in [20] that in the limit when the number of particles in the heat bath tends to infinity, the motion of the Brownian particle is governed (exactly!) by the stochastic differential equations,

$$\dot{q}(t) = p(t) \tag{1}$$

$$dp(t) = -ap(t)\mathrm{d}t + k\mathrm{d}w(t)\,, \quad a > 0 \tag{2}$$

the second of which is the ubiquitous *Langevin equation* of statistical physics. This equation contains a dynamical *friction term*, $(-ap(t))$, and a forcing function term $w(t)$ which is described mathematically as a Wiener process. Physically, these terms can be interpreted as the influence of the surrounding medium on the observed particle, the friction accounting for the energy transferred from the particle to the heat bath and the forcing term $w(t)$ representing the sum of infinitesimal "random shocks" inflicted to the particle by the surrounding medium.

The stochastic process describing the temporal evolution of the observables $q(t)$ and $p(t)$ is a Markov diffusion process (in fact, both $p(t)$ and the two-dimensional vector process $[q(t)\ p(t)]$ are Markov). The evolution of the macroscopic observables is hence *dissipative*, because of the friction term, and *irreversible*, since diffusion processes evolve in time by the action of a semigroup which cannot be inverted to become a group. This is in accordance with the basic postulates of (macroscopic) thermodynamics.

The Ford–Kac–Mazur example can be generalized [48,49,50,52]. Assume a microscopic system described by a Hamiltonian H on the phase space Ω, a smooth manifold of very large dimension $2N$. The solution of the Hamilton equations determines the phase of the system, say configurations and momenta $z(t) := [q(t)\ \ p(t)]'$ at each time t, uniquely in terms of the initial value $z(0) \in \Omega$. The correspondence defines a flow $z(t) = \Phi(t)z(0)$ on the phase space which leaves invariant the total energy, $H(z(t))$, of the system. The microscopic evolution is in this sense, reversible and conservative. However, microscopic phenomena are *complex* in the sense that they involve an enormous number of interacting components and it is impossible to keep track of the dynamics of such a complex system. One must then resort to a statistical description.

A probability distribution on Ω which is *invariant for the Hamiltonian flow* $U(t)$ defines "thermal equilibrium". It is known that in a finite-dimensional space any absolutely continuous $U(t)$-invariant probability measure admits a one parameter family of densities $\rho(z)$ of the Maxwell–Boltzmann type, equal to a normalization constant times $\exp[-\frac{1}{2\beta}H(z)]$, $\beta > 0$, being interpreted as the "absolute temperature". For a quadratic Hamiltonian in a linear phase space these distributions are Gaussian. The choice of an initial phase $z(0)$ of a system in thermal equilibrium can then be interpreted as a random choice of an elementary event in the elementary outcome space $\{\Omega,\ \rho\}$.

It follows that in thermal equilibrium any observable; i. e. any measurable function $h\colon \Omega \to \mathbb{R}$ on the phase space can be regarded as a random variable. In fact, once combined with the (measure-preserving) Hamiltonian flow $z(t) = U(t)z(0)$, any measurable observable defines a *stationary stochastic process*. In general, the time evolution of any finite, say m-dimensional, family of observables $\{h_1, \dots, h_m\}$ can be identified with a vector-valued m-dimensional stationary process $\{y(t)\}$ on Ω, with components

$$y_k(t, z(0)) := h_k(U(t)z(0)) \tag{3}$$

where $z(0) \in \Omega$ is the elementary event and $U(t)$ the measure preserving group of transformations on the underlying probability space.

The main point of this story is that the "randomization" of the phase space may, under certain conditions, lead to a description of certain observable processes $\{y(t)\}$ of (3) by a stochastic dynamical model of a *much simpler structure than the microscopic deterministic Hamiltonian description*. These could for example be representations of $\{y(t)\}$ of the type $y(t) = h(x(t))$ where $\{x(t)\}$ a finite dimensional, say $\mathbb{R}^n$-valued Markov process and h is a suitable function $h\colon \mathbb{R}^n \to \mathbb{R}^m$. This description, in case the Markovian representation has a smaller complexity (dimension) than that of the microscopic description, would precisely meet both the general physical plausibility for a thermodynamic description and the mathematical requirement of reduction of complexity.

Remark 1 This view generalizes the way of conceiving stochastic aggregation in the physical literature, e.g [31,32], where it is generally required that the observables themselves should be Markov and satisfy a Langevin equation.

In general the requirement that the observables should themselves be Markov processes is seldom a reasonable one. For example, in the Brownian particle example the macroscopic observable $q(t)$ by itself need *not* satisfy any stochastic differential equation. Formally, $q(t)$ is just a memoryless function of a two-dimensional Markov process $x(t) := [q(t)\ p(t)]'$ and it is instead the "thermodynamic state" $x(t)$ which satisfies a stochastic differential equation.

Stochastic Realization

Stochastic realization is the abstract version of the aggregation problem mentioned above. One studies the problem of representing an m-dimensional stationary process $\{y(t)\}$ as a *memoryless function of a Markov process*. In this setting the physical phase space and the Hamiltonian group $\{U_t\}$ are generalized to an abstract probability space and to a measure preserving one-parameter group of transformations mapping Ω onto itself (the shift group of the process). In this article we shall assume that $\{y(t)\}$ is smooth with continuous values but the problem area also concerns stationary finite state-processes, see [5,19,46,66], which have important applications in communications and biology. In the present context, the question is when does a given stationary, m-dimensional, stochastic process $\{y(t)\}$ admit representations of the type $y(t) = h(x(t))$ where $\{x(t)\}$ is a Markov diffusion process taking values on a finite-dimensional state space X and h is say a continuous function from X to $\mathbb{R}^m$.

Since a smooth Markov diffusion process $\{x(t)\}$ is a solution of a stochastic differential equation (Chap. VI, Par. 3 in [14]), any such process must admit representations as the output of a stochastic dynamical system of the type

$$\mathrm{d}x(t) = f(x(t))\mathrm{d}t + G(x(t))\mathrm{d}w(t) \tag{4}$$

$$y(t) = h(x(t)) \tag{5}$$

where $\{w(t)\}$ is a vector m Wiener process. This stochastic dynamical system generalizes the Langevin equation representation. For a Gaussian mean square continuous stationary process, the representation simplifies to a linear model of the form

$$\mathrm{d}x(t) = Ax(t)\mathrm{d}t + B\mathrm{d}w(t) \tag{6}$$

$$y(t) = Cx(t) \tag{7}$$

where A, B, C are constant matrices of appropriate dimensions. There may be a need for introducing an additive noise term also in the second (output) equation when the process y has stationary increments, a situation more general than stationarity, see [37].

Representations as the output of a stochastic dynamical system are called *stochastic realizations* or *state-space realizations* of the process y and the Markov process $x(t)$ is accordingly called the *state process* of y, or of the realization.

Signal descriptions by a finite dimensional stochastic realization can be transformed into other equivalent forms such as ARMA models etc., which are also widely used in the engineering and econometric literature. It should be stressed that, irrespective of which particular equivalent form of the stochastic model, virtually all sequential signal processing, prediction and control algorithms (Kalman filtering to name just the most popular) require availability of a finite-dimensional realization of one form or another. Hence signal representation by a stochastic system is indeed a crucial prerequisite for system analysis and design in engineering and applied sciences.

A substantial amount of literature on the stochastic realization problem has appeared in the last four decades. The stationary (and stationary increments) Gaussian case is now covered by a rather complete linear theory, and is surveyed e. g. in [35,37]. For a more up to date account, see the forthcoming book [41]. The nonlinear realization problem is still underdeveloped, see [45,51,62].

Wide-Sense Stochastic Realization

We shall discuss only the wide-sense version of the realization problem, where random variables and processes are described in terms of first- and second-order moments. Since first- and second-order moments individuate a Gaussian distribution uniquely, the theory will in particular cover the Gaussian case. Hereafter, wide-sense stationarity will be simply referred to as stationarity. Also, a "wide-sense Markov process" will simply be called a "Markov process".

The wide-sense realization problem may look like a very particular modeling problem to deal with, but it is completely solvable and is general enough to reveal quite explicitly some of the important features of the solution set of stochastic realizations, some of which are quite unexpected. It is also motivated by its wide applicability since most times in practice the only reasonable way to describe a priori random phenomena is by second-order moments, say by covariances or spectra.

We advise the reader that in many applied fields one may want to construct dynamical system models involving also exogenous input (or decision) variables. The stochastic realization problem of constructing linear state-space representations of a stationary process with inputs has been studied in [11,53,54,55]. In the last two references one may find also material related to the germane problem of *identification with inputs*.

For reasons of space limitations we shall not touch upon this subject. A pointer to the relevant identification literature will be given in the final section of the article.

We shall discuss in some detail only stochastic realization of *discrete-time* random processes. An analogous (and in fact somewhat simpler) treatment can be given for continuous time processes [35] but we choose discrete-time since this theory makes contact with system identification which we shall discuss later in the section entitled "Dynamical System Identification".

Given a vector (say m-dimensional) wide-sense stationary purely nondeterministic (purely nondeterministic will be abbreviated to p.n.d. in the following; this property is called *linear regularity* in the Russian literature, see [57]) process $y = \{y(t)\}$ with $t \in \mathbb{Z}$ (the integers), one wants to find representations of y in terms of a finite dimensional Markov process. In other words one wants to find linear representations of the stationary process y, of the form

$$x(t+1) = Ax(t) + Bw(t) \tag{8a}$$

$$y(t) = Cx(t) + Dw(t) \tag{8b}$$

and procedures for constructing them. Here $\{w(t)\}$ is a vector, p-dimensional normalized white noise process, i. e. $\mathbb{E}\{w(t)w(s)'\} = I\delta(t-s)$, $\mathbb{E}\{w(t)\} = 0$, δ being the Kronecker delta function. Note that a white noise w could be seen as a (degenerate) kind of Markov process. The reason for having the term $Dw(t)$ in (8b) is to avoid the presence of such degenerate components in the dynamic equation for the state process x which would artificially increase its dimension. With this convention, in the extreme circumstance where $y = w$, the state dimension (of a minimal realization) is zero.

The linear representation (8a) involves auxiliary variables, i. e. random quantities which are not given as a part of the original data, especially the n-dimensional state process x (a stationary Markov process) and the generating white noise w. The peculiar properties of these processes actually embody the desired system structure (8a). Constructing these auxiliary processes is an essential part of the realization problem.

We shall restrict ourselves to representations (8a) for which

- (A,B,C) is a minimal triplet; i. e. (A, B) is a reachable pair and (A, C) is an observable pair, see e. g. [24] for these notions;
- $\begin{bmatrix} B \\ D \end{bmatrix}$ has independent columns.

These are classical conditions of nonredundancy of the model [24] and entail no loss of generality. From standard spectral representation theory, see e. g., Chap. 1 in [57], it follows that Eq. (8a) admits stationary solutions if and only if the rows of the $n \times p$ matrix $(e^{i\theta}I - A)^{-1}B$ are square integrable functions of $\theta \in [-\pi, \pi]$. This is equivalent to the absence of poles of the function $z \to (zI - A)^{-1}B$ on the unit circle. Equivalently, the eigenvalues of A of modulus one, if any, must be "unreachable" for the pair (A, B).

When eigenvalues on the unit circle are present, to guarantee stationarity they must be simple roots of the minimal polynomial of A. These eigenvalues, necessarily in even number say $2k$, give then rise to a sum of k uncorrelated sinusoidal oscillations with random amplitude, the so-called *purely deterministic* component of the process. This purely deterministic component of the stationary Markov process x obeys a fixed undriven (i. e. deterministic) linear difference equation of the type $x(t+1) = A_o x(t)$, which is the restriction of (8a) to the modulus one eigenspace of A. The initial conditions are random variables independent of the driving noise w. Since there is no stochastic forcing term in this restricted state equation, it follows that the presence of a purely deterministic component in the Markov process necessarily implies that the pair (A, B) must be nonreachable. The first of the two conditions of nonredundancy listed above hence excludes the presence of a purely deterministic component in the state, and therefore also in the output process y.

From this discussion it is immediately seen that there are stationary solutions x of the difference Eq. (8a) which are p.n.d. Markov processes if and only if A does not have eigenvalues of modulus one; i. e.

$$|\lambda(A)| \neq 1 \,. \tag{9}$$

When $|\lambda(A)| < 1$, the representation (8a) is called *forward* or *causal*. When $|\lambda(A)| > 1$ the representation is called *backward* or *anticausal*. Traditionally the causality of a representation, i. e. the condition $|\lambda(A)| < 1$ has been regarded as being equivalent to stationarity. In fact, stationarity permits a whole family of representations (8a) of the (same) process x, where the matrices A can have very different spectra. Formulas for computing a backward representation starting from a forward one or vice versa are given in [34,36]. The two representations (in particular the relative A matrices) are related by a particular family of *linear state feedback* transformations [47].

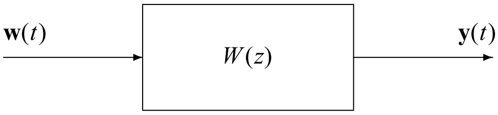
The *covariance function* of a zero-mean wide sense stationary process y is the $m \times m$ matrix function $k \mapsto \Lambda_k$, defined by

$$\Lambda_k := \mathbb{E}\{y(t+k)y(t)'\} = \mathbb{E}\{y(k)y(0)'\} \qquad k \in \mathbb{Z} \,.$$

This matrix function will be the initial data of our problem. Since y is p.n.d. the function Λ admits a Fourier transform called the *spectral density* matrix of y, given by

$$\Phi(z) = \sum_{t=-\infty}^{\infty} \Lambda_t z^{-t}$$

where $z = e^{i\theta}$; $\theta \in [-\pi, \pi]$. It is easy to see that the spectral density matrix of a process admitting a state-space realization (8a) must be a rational function of z. This fact follows easily by eliminating the variable x in (8a) and thereby expressing y as the output of a linear filter as in Fig. 1 below whose transfer function $W(z) = C(zI - A)^{-1}B + D$ is a rational function of z, and then using the classical Kintchine and Wiener formula for the spectrum of a filtered



Stochastic Noises, Observation, Identification and Realization with, Figure 1
Shaping filter representation of a stochastic process

stationary process [28,69] so that,

$$\Phi(z) = W(z)W(1/z)' \,. \tag{10}$$

Hence we have

Proposition 1 *The transfer function $W(z) = C(zI - A)^{-1}B + D$ of any state space representation (8a) of the stationary process y, is a* spectral factor *of Φ, in the sense that it satisfies the spectral factorization equation (10).*

It is easy and instructive to check this directly. Supposing that one has a *causal* realization (8a), one can compute for $k \geq 0$,

$$\begin{aligned}\Lambda_k &= \mathbb{E}[\, CA^k x(t) \\ &+ \sum_{s=0}^{k-1} CA^{k-1-s}Bw(t+s) + Dw(t+k)\,] \\ &\qquad\qquad [Cx(t) + Dw(t)]' \end{aligned} \tag{11}$$

$$= CA^{k-1}\bar{C}' + \frac{1}{2}\Lambda_0\delta(t) \,, \tag{12}$$

where $\bar{C}' = \mathbb{E}x(t+1)y(t)'$ and $\Lambda_0 = \mathbb{E}y(t)y(t)'$. By identifying terms, these matrices can be expressed directly in terms of the realization parameters A, B, C, D, by

$$P = APA' + BB' \tag{13a}$$

$$\bar{C}' = APC' + BD' \tag{13b}$$

$$\Lambda_0 = CPC' + DD' \tag{13c}$$

the matrix $P = \mathbb{E}x(t)x(t)'$ being the covariance matrix of the state process. Note that P obeys Eq. (13a) since by causality ($|\lambda(A)| < 1$) $x(t)$ is a function of the infinite past $\{w(s);\ s < t\}$ and hence is uncorrelated with $w(t)$ so that the two terms in the right hand side of (8a) are also uncorrelated and their variances add. In fact, by reachability of the model, P is actually (symmetric and) positive definite.

Now, introduce the decomposition

$$\Phi(z) = \Phi_+(z) + \Phi_+(1/z)' \tag{14}$$

where $\Phi_+(z) = \frac{1}{2}\Lambda_0 + \Lambda_1 z^{-1} + \Lambda_2 z^{-2} + \cdots$ is analytic outside of the unit circle. Clearly $\Phi_+(z)$ must be the Fourier transform of the "causal" tract of the covariance function Λ, computed in (12). Hence

$$\Phi_+(z) = C(zI - A)^{-1}\bar{C}' + \frac{1}{2}\Lambda_0 \,. \tag{15}$$

This shows that the spectrum of a process y described by a state-space model (8a), besides being a rational function of z, is directly expressible in parametric form in terms of the parameters of a causal realization. This explicit computation of the spectrum seems to be due to R.E. Kalman and B.D.O. Anderson [4,21].

"Classical" stochastic realization theory, developed in the late 1960s [3,4,17], deals with the inverse problem of computing the parameters A, B, C, D of a state space realization starting from the spectrum (or, equivalently, of the covariance function) of a stationary process y. Of course one assumes here that y is p.n.d. and has a spectral density Φ which is a rational function of $z = e^{j\theta}$.

This inverse problem is just a parametric version of the spectral factorization problem: Given a rational spectrum $\Phi(z)$ one looks for rational spectral factors parametrized as $W(z) = C(zI - A)^{-1}B + D$, solutions of (10). Since rational factors of a given spectral density matrix are in general infinitely many, one must preliminarily single out the interesting solutions by imposing certain restrictions. One very reasonable restriction which is imposed on the solutions of (10) is that the spectral factors should be of *minimal degree*. The degree of a proper rational matrix function is by definition the dimension n of the state of a minimal realization. Minimal degree (for short, *minimal*) spectral factors must have a degree which is exactly one half of the degree of the spectral density matrix $\Phi(z)$.

The solution of the inverse problem is described in the following proposition.

Proposition 2 *Assume the spectrum is given in the parametric form (15) where $|\lambda(A)| < 1$ and $(A, C, \bar{C}')$ is a minimal triplet of dimension n. Then the minimal degree spectral factors of $\Phi(z)$,* analytic in$\{|z| > 1\}$, *are in one-to-one correspondence with the symmetric $n \times n$ matrices P solving the* Linear Matrix Inequality *(LMI)*

$$M(P) := \begin{bmatrix} P - APA' & \bar{C}' - APC' \\ \bar{C} - CPA' & \Lambda_0 - CPC' \end{bmatrix} \geq 0 \tag{16}$$

where the symbol $\geq$ means positive semidefinite. The correspondence is to be understood in the following sense:

Corresponding to each solution $P = P'$ of (16), consider the full column rank matrix factor $\begin{bmatrix} B \\ D \end{bmatrix}$ of $M(P)$,

$$M(P) = \begin{bmatrix} B \\ D \end{bmatrix} [B'D'] \tag{17}$$

(this factor is unique modulo right multiplication by orthogonal matrices) and form the rational matrix

$$W(z) := C(zI - A)^{-1}B + D \,. \tag{18}$$

Then (18) is a minimal realization of a minimal analytic spectral factor of $\Phi(z)$ and all minimal analytic factors can be obtained in this way.

Note that any matrix P solving the LMI satisfies the system (12) and hence can be interpreted as the state covariance of a causal realization of y. It follows that any P solving the LMI is necessarily positive definite.

In fact it has been shown [17] that the set of symmetric solutions to the LMI (16)

$$\mathcal{P} := \{P = P' \, ; \, M(P) \geq 0\}$$

is a closed, bounded and convex set with maximal and minimal elements $P_-, \, P_+ \in \mathcal{P}$ which satisfy

$$P_- \leq P \leq P_+ \quad \text{for all} \ \ P \in \mathcal{P}$$

where $P_1 \leq P_2$ means that $P_2 - P_1 \geq 0$ (is positive semidefinite).

The existence of solutions to the LMI is equivalent to the property of $\Phi_+(z)$ defined in (14) of being a *positive-real matrix function*, namely that Φ_+ should be analytic in $\{|z| > 1\}$ and that $\Re e \, \Phi_+(e^{j\theta}) \geq 0$. The second condition is automatically satisfied when Φ admits spectral factors since

$$2\Re e \, \Phi_+(e^{j\theta}) = W(e^{j\theta}) W(e^{j\theta})^* \geq 0$$

(the star meaning complex conjugate transpose). This property can be characterized in general, for not necessarily stable representations, by the so-called *positive-real lemma* discovered in the 1960s by Kalman, Yakubovich, and Popov [22,56,71] in the context of dissipative systems and stability theory. The first application of the Kalman–Yakubovich–Popov theory to factorization of spectral density functions and to stochastic realization (called *stationary covariance generation*) is due to B.D.O. Anderson [3,4]. Far reaching generalizations to not necessarily stable systems have been pursued by J.C. Willems in his widely quoted paper [70]. The terminology "positive real" derives from network synthesis.

Under certain conditions on the zeros of the spectrum, discussed e. g. in [18], the matrix $\Delta(P) := \Lambda_0 - CPC'$ is nonsingular (and hence positive definite) for all $P \in \mathcal{P}$. In this case (16) is equivalent to the nonnegativity of the Schur complement of $\Delta(P)$ in the matrix $M(P)$, which can be written

$$P - APA' - (\bar{C}' - APC')\Delta(P)^{-1}(\bar{C}' - APC')' \leq 0 \, . \quad (19)$$

This is the *Algebraic Riccati Inequality* (ARI) of spectral factorization, first analyzed in the work of [3,17]. The solutions of the ARI with equality sign; i. e. the solutions of the Algebraic Riccati Equation (ARE)

$$P - APA' - (\bar{C}' - APC')\Delta(P)^{-1}(\bar{C}' - APC')' = 0 \quad (20)$$

make $M(P)$ of minimal rank $m = \dim y$ and hence correspond to *square spectral factors* (18). These spectral factors all have the same denominator matrix A and hence only differ by the location of their zeros. In particular one may show that $P_-, \, P_+$ solve the ARE and the two spectral factors W_- and W_+ corresponding to $P_-, \, P_+$ (are square and) have all of their zeros inside and respectively outside of the unit circle. In other words, W_- is the *minimum phase* and W_+ the *maximum phase* spectral factor. These spectral factors are essentially unique. All other square spectral factors are obtained (in rough terms) by "flipping" some of the zeros of W_- to their reciprocals outside of the unit circle. The maximum phase factor, W_+, is obtained by flipping *all* of the zeros of W_- to their reciprocals. More information on the zero structure of the minimal spectral factors can be found in [39].

In the next section we shall analyze in more detail the correspondence between spectral factors and stochastic realizations. As we shall see, for square spectral factors this correspondence is one-to-one and hence this particular class of stochastic realizations, for which $\dim w = \dim y = m$ can be classified according to their zero location.

The LMI plays an important role in many areas of system theory such as stability theory, dissipative systems and is central in H^∞ control and estimation theory. It seems to be much less appreciated in the scientific community that it plays a very basic role in modeling of stationary random signals as well. As we shall see in the next section certain solutions of the LMI (or of the ARI) have special probabilistic properties and are related to Kalman-filter or "innovations-type" realizations.

Geometric Stochastic Realization

The classical stationary covariance realization theory of the previous section is purely distributional as it says nothing about representation of random quantities in a truly probabilistic sense (i. e. how to generate the random variables or the sample paths of a given process, not just its covariance function). This was implicitly pointed out by Kalman already in [23]. In the last decades a *geometric coordinate-free* approach to stochastic modeling has been put forward in a series of papers by Lindquist, Picci, Ruckebusch et al. [33,34,35,58,59,60] which aims at the representation of random processes in this more specific sense. This motivation is also present in the early papers by Akaike [1,2].

A main point of the geometric approach is to provide procedures for the construction of the random quantities defining a stochastic state-space realization. In particular the *state space* of a realization is defined in terms of the conditional independence relation between past and future of the signals involved. This relation is intrinsically *coordinate-free* and in the present setting involves only linear subspaces of a given ambient Hilbert space of random variables, typically made of linear functionals of the variables of the process y to be modeled (but in some situations other random data may be used to construct the model).

Constructing the State Space of a Stationary Process

The theme of this section will be a review of geometric realization theory for the stochastic process y. The geometric theory centers on the idea of *Markovian Splitting Subspace* for the process y. This concept is the probabilistic analog of the deterministic notion of state space of a dynamical system and captures at an abstract level the property of "dynamic memory" that the state variables have in deterministic dynamical system theory. Once a stochastic state space is given the construction of the auxiliary random quantities which enter in a state space model and in particular the state process is fairly obvious. The state vector $x(t)$ of a particular realization can be regarded just as a particular basis for the state space, hence once a state-space is constructed, finding state equations is just a matter of choosing a basis and computing coordinates.

Notation 1 *In what follows the symbols* $\vee$, $+$ *and* $\oplus$ *will denote vector sum,* direct *vector sum and* orthogonal *vector sum of subspaces, the symbol* $\mathbf{x}^{\perp}$ *will denote the orthogonal complement of a (closed) subspace* $\mathbf{X}$ *of a Hilbert space with respect to some predefined ambient space. Given a collection* $\{X_\alpha \mid \alpha \in A\}$ *of subsets of a Hilbert space* $\mathbf{H}$ *we shall denote by* $\overline{\text{span}}\{X_\alpha \mid \alpha \in A\}$ *the closure in* $\mathbf{H}$ *of the linear (real) vector space generated by the collection. The orthogonal projection onto the subspace* $\mathbf{X}$ *will be denoted by the symbol* $\mathbb{E}(\cdot \mid \mathbf{X})$ *or by the shorthand* $\mathbb{E}^{\mathbf{X}}$*. The notation* $\mathbb{E}(z \mid \mathbf{X})$ *will be used also when* z *is vector-valued. The symbol will then denote the vector with components* $E(z_k \mid \mathbf{X})$, $k = 1, \ldots$*. For vector quantities,* $|v|$ *will denote Euclidean length (or absolute value in the scalar case).*

Let y be a zero mean stationary vector process with finite second-order moments defined on some underlying probability space $\{\Omega, \mathcal{A}, \mu\}$ and let $L^2\{\Omega, \mathcal{A}, \mu\}$ denote the Hilbert space of second-order random variables defined on Ω, endowed with the inner product $< \xi, \eta >= \mathbb{E}\{\xi\eta\}$. Let $\mathbf{H}(y)$ be the (closed) linear subspace of $L^2\{\Omega, \mathcal{A}, \mu\}$, linearly generated by the variables of the process; i. e.

$$\mathbf{H}(y) := \overline{\text{span}}\{y_k(t)\,; k = 1, 2, \ldots, m;\ t \in \mathbb{Z}\}\,.$$

If y is generated by a linear model of the type (8a) this Hilbert space will in general be a proper subspace of the space $\mathbf{H}(w)$, generated by the input white noise of the model (8a). All random variables of the stochastic system (8a) belong to $\mathbf{H}(w)$ and for this reason $\mathbf{H}(w)$ is called the *ambient space* of the model. By stationarity of w, $\mathbf{H}(w)$ comes naturally equipped with a unitary operator U, called the *shift* of w, such that $Uw_i(t) = w_i(t+1)$ for $i = 1, 2, \ldots, p$ and all $t \in \mathbb{Z}$ (stationarity). Note that the family $\left\{U^t; t \in \mathbb{Z}\right\}$ forms a one parameter group of unitary operators on the ambient space; in particular, $U^{-t} = (U^*)^t$. The pair $(\mathbf{H}(w), U)$ is also called a *stationary Hilbert space*. Clearly the processes x and y will also be stationary with respect to U.

The *past* and *future* subspaces (at time zero) of a stationary process are

$$\mathbf{H}^-(y) := \overline{\text{span}}\{y_k(t)\,; k = 1, 2, \ldots, m;\ t < 0\}$$
$$\mathbf{H}^+(y) := \overline{\text{span}}\{y_k(t)\,; k = 1, 2, \ldots, m;\ t \geq 0\}\,.$$

Because of stationarity everything propagates in time by the action of the unitary group, e. g. the past and future at time t are obtained as $\mathbf{H}_t^-(y) = U^t\mathbf{H}^-(y)$ and $\mathbf{H}_t^+(y) = U^t\mathbf{H}^+(y)$. For this reason the geometric definitions and constructions to follow, although valid for an arbitrary time instant will usually be referred to the time instant $t = 0$.

Let $\mathbf{X}$ be a subspace of some large stationary Hilbert space $\mathbf{H}$ of wide-sense random variables containing $\mathbf{H}(y)$. Define

$$\mathbf{X}_t := U^t\mathbf{X}, \quad \mathbf{X}_t^- := \vee_{s \leq t}\mathbf{X}_s, \quad \mathbf{X}_t^+ := \vee_{s \geq t}\mathbf{X}_s.$$

Definition 1 A *Markovian Splitting Subspace* $\mathbf{X}$ for the process y is a subspace of $\mathbf{H}$ making the vector sums $\mathbf{H}(y)^- \vee \mathbf{X}^-$ and $\mathbf{H}(y)^+ \vee \mathbf{X}^+$ conditionally orthogonal (i. e. conditionally uncorrelated) given $\mathbf{X}$, denoted,

$$\mathbf{H}(y)^- \vee \mathbf{X}^- \perp \mathbf{H}(y)^+ \vee \mathbf{X}^+ \mid \mathbf{X}. \tag{21}$$

The conditional orthogonality condition (21) can be equivalently written as

$$\mathbb{E}[\mathbf{H}(y)^+ \vee \mathbf{X}^+ \mid \mathbf{H}(y)^- \vee \mathbf{X}^-] = \mathbb{E}[\mathbf{H}(y)^+ \vee \mathbf{X}^+ \mid \mathbf{X}] \tag{22}$$

$$\mathbb{E}[\mathbf{H}(y)^- \vee \mathbf{X}^- \mid \mathbf{H}(y)^+ \vee \mathbf{X}^+] = \mathbb{E}[\mathbf{H}(y)^- \vee \mathbf{X}^- \mid \mathbf{X}] \tag{23}$$

which formalize the intuitive meaning of the splitting subspace $\mathbf{X}$ as a dynamic memory of the past (future) for the purpose of predicting the joint future (past). It follows in particular that $\mathbb{X}$ is Markovian (i. e. $\mathbf{X}^- \perp \mathbf{X}^+ \mid \mathbf{X}$) and any basis vector $x := [x_1, x_2, \ldots, x_n]'$ in a (finite-dimensional) Markovian splitting subspace $\mathbf{X}$ generates a stationary Markov process $x(t) := U^t x$, $t \in \mathbb{Z}$ which, as we shall see in a moment, serves as a *state* for the process y.

The subspace $\mathbf{X}$ is called *proper*, or *p.n.d.* if

$$\cap_t \mathbf{Y}_t^- \vee \mathbf{X}_t^- = \{0\}, \text{ and } \cap_t \mathbf{H}(y)_t^+ \vee \mathbf{X}_t^+ = \{0\}\,.$$

Obviously for the existence of proper splitting subspaces y must also be purely nondeterministic [57]. Properness is, by the Wold decomposition theorem, equivalent to the existence of two vector white noise processes w and $\bar{w}$ such that,

$$\mathbf{H}(y)^- \vee \mathbf{X}^- = \mathbf{H}^-(w)\,, \quad \mathbf{H}(y)^+ \vee \mathbf{X}^+ = \mathbf{H}^+(\bar{w})\,.$$

If $\mathbf{X}$ is proper, every Markov process $x(t) := U^t x$ is purely nondeterministic.

The subspaces

$$\mathbf{S} := \mathbf{H}(y)^- \vee \mathbf{X}^- \quad \text{and} \quad \bar{\mathbf{S}} := \mathbf{Y}^+ \vee \mathbf{X}^+ \tag{24}$$

associated to a Markovian Splitting subspace $\mathbf{X}$, play an important role in the geometric theory of stochastic systems. They are called the *scattering pair* of $\mathbf{X}$ as they can be seen to form an incoming–outgoing pair in the sense of Lax–Phillips Scattering Theory [29].

Definition 2 Given a stationary Hilbert space $(\mathbf{H}, U)$ containing $\mathbf{H}(y)$, a *scattering pair* for the process y is a pair of subspaces $(\mathbf{S}, \bar{\mathbf{S}})$ satisfying the following conditions for a stationary process,

1. $U^*\mathbf{S} \subset \mathbf{S}$ and $U\bar{\mathbf{S}} \subset \bar{\mathbf{S}}$, i. e. $\mathbf{S}$ and $\bar{\mathbf{S}}$ are invariant for the left and right shift semigroups (this means that $\mathbf{S}_t$ is increasing and $\bar{\mathbf{S}}_t$ is decreasing with time).
2. $\mathbf{S} \vee \bar{\mathbf{S}} = \mathbf{H}$
3. $\mathbf{S} \supset \mathbf{H}(y)^-$ and $\bar{\mathbf{S}} \supset \mathbf{H}(y)^+$
4. $\mathbf{S}^\perp \subset \bar{\mathbf{S}}$ or, equivalently, $\bar{\mathbf{S}}^\perp \subset \mathbf{S}$

The following representation Theorem provides the link between Markovian splitting subspaces and scattering pairs.

Theorem 1 ([35]) *The intersection*

$$\mathbf{X} = \mathbf{S} \cap \bar{\mathbf{S}} \tag{25}$$

of any scattering pair of subspaces of $\mathbf{H}$ *is a Markovian splitting subspace. Conversely every Markovian splitting subspace can be represented as the intersection of a scattering pair. The correspondence* $\mathbf{X} \leftrightarrow (\mathbf{S}, \bar{\mathbf{S}})$ *is one-to-one, the scattering pair corresponding to* $\mathbf{X}$ *being given by*

$$\mathbf{S} = \mathbf{H}(y)^- \vee \mathbf{X}^- \qquad \bar{\mathbf{S}} = \mathbf{H}(y)^+ \vee \mathbf{X}^+\,. \tag{26}$$

The process of forming scattering pairs associated to $\mathbf{X}$ should be thought of as an "extension" of the past and future spaces of y. The rationale for this extension is that scattering pairs have an extremely simple splitting geometry due to the fact that

$$\mathbf{S} \perp \bar{\mathbf{S}} \mid \mathbf{S} \cap \bar{\mathbf{S}} \tag{27}$$

equivalent to

$$\mathbf{S} \vee \bar{\mathbf{S}} = \bar{\mathbf{S}}^\perp \oplus (\mathbf{S} \cap \bar{\mathbf{S}}) \oplus \mathbf{S}^\perp \tag{28}$$

which is called *perpendicular intersection*. It is easy to show that Property (4) in the definition of a scattering pair is actually equivalent to perpendicular intersection. This property of conditional orthogonality given the intersection can also be seen as a natural generalization of the Markov property. Indeed for a Markovian family of subspaces $\mathbf{X} = \mathbf{X}^- \cap \mathbf{X}^+$ and $\mathbf{S} = \mathbf{X}^-$, $\bar{\mathbf{S}} = \mathbf{X}^+$ intersect perpendicularly with intersection $\mathbf{X}$.

Note that $\mathbf{A} \perp \mathbf{B} \mid \mathbf{X} \Rightarrow \mathbf{A} \cap \mathbf{B} \subset \mathbf{X}$ but the inclusion of the intersection in the splitting subspace $\mathbf{X}$ is only *proper* in general. For perpendicularly intersecting subspaces, the intersection is actually the *unique minimal subspace* making them conditionally orthogonal.

Theorem 1 is the fundamental device for the construction and classification of Markovian splitting subspaces.

Denote by $\mathbf{W}_t, \bar{\mathbf{W}}_t$ the (wandering) subspaces spanned by the components, at time t, of the generating noises $w(t)$ and $\bar{w}(t)$, of the scattering pair of $\mathbf{X}$. Since

$$\mathbf{S}_{t+1} = \mathbf{S}_t \oplus \mathbf{W}_t\,, \tag{29}$$

we can write,

$$\mathbf{X}_{t+1} = \mathbf{S}_{t+1} \cap \bar{\mathbf{S}}_{t+1} \subset (\mathbf{S}_t \cap \bar{\mathbf{S}}_{t+1}) \oplus (\mathbf{W}_t \cap \bar{\mathbf{S}}_{t+1}) \tag{30}$$

Since $\bar{\mathbf{S}}_t$ is decreasing in time, we have $\mathbf{S}_t \cap \bar{\mathbf{S}}_{t+1} \subset \mathbf{X}_t$ and by projecting the shifted basis $Ux(t) := x(t+1)$, onto the last orthogonal direct sum above, the time evolution of any basis vector $x(t) := [x_1(t), x_2(t), \ldots, x_n(t)]'$ in $\mathbf{X}_t$, is described by a linear equation of the type

$$x(t+1) = Ax(t) + Bw(t)\,.$$

It is also easy to see that, by the p.n.d. property, A must have all its eigenvalues strictly inside of the unit circle. Naturally, by decomposing instead $\bar{\mathbf{S}}_{t-1} = \bar{\mathbf{S}}_t \oplus \bar{\mathbf{W}}_t$ one could have obtained a *backward difference equation* model for the Markov process x, driven by the backward generating noise $\bar{w}$.

To complete the representation, note that by definition of the past space, $y(t) \in (\mathbf{S}_{t+1} \cap \bar{\mathbf{S}}_t)$. Inserting the decomposition (29) and projecting $y(t)$, leads to a state-output equation of the form

$$y(t) = Cx(t) + Dw(t) .$$

Here one could also obtain a state-output equation driven by the backward noise $\bar{w}$, the same noise driving the backward state model obtained before.

As we have just seen, any basis in a Markovian splitting subspaces produces a stochastic realization of y. It is easy to reverse the implication. In fact the following fundamental characterization holds.

Theorem 2 ([33,37]) *The state space* $\mathbf{X} = \text{span}\{x_1(0), x_2(0), \dots, x_n(0)\}$ *of any stochastic realization (8a) is a Markovian splitting subspace for the process y.*

Conversely, given a finite-dimensional Markovian splitting subspace $\mathbf{X}$*, to any choice of basis* $x(0) = [\, x_1(0), x_2(0), \dots, x_n(0)\,]'$ *in* $\mathbf{X}$ *there corresponds a stochastic realization of y of the type (8a).*

Once a basis in $\mathbf{X}$ is available, there are obvious formulas to compute the parameters $(A, C, \bar{C})$ of the corresponding stochastic realization, namely

$$A = \mathbb{E}x(t+1)x(t)'\, P^{-1} \tag{31}$$

$$C = \mathbb{E}y(t)x(t)'\, P^{-1} \tag{32}$$

$$\bar{C} = \mathbb{E}y(t-1)x(t)' \tag{33}$$

where $P = \mathbb{E}x(t)x(t)'$ is the state covariance matrix (i. e. the Gramian matrix of the basis). The matrices B and D, however, are related to the (unobservable) generating white noise w and require the solution of the LMI. This abstract procedure which permits us to compute the parameters of a stochastic realization once the state has been constructed, can be rendered quite concrete and forms the conceptual basis of subspace identification.

Stochastic realizations are called *internal* when $\mathbf{H} = \mathbf{H}(y)$, i. e. the state space is built from the Hilbert space made just of the linear statistics of the process y. For identification the only realizations of interest are the internal ones.

A central problem of geometric realization theory is to construct and to classify all minimal state spaces, i. e. the minimal Markovian splitting subspaces for the process y.

The obvious ordering of subspaces of $\mathbf{H}$ by inclusion, induces an ordering on the family of Markovian splitting subspaces. The notion of minimality is most naturally defined with respect to this ordering. Note that this definition is independent of assumptions of finite-dimensionality and applies also to infinite-dimensional Markovian splitting subspaces, i. e. to situations where comparing dimension would not make much sense.

Definition 3 A Markovian splitting subspace is *minimal* if it does not contain (properly) other Markovian splitting subspaces.

The study of minimality forms an elegant chapter of stochastic system theory. There are several known geometric and algebraic characterizations of minimality of splitting subspaces and of the corresponding stochastic state-space realizations. Since, however, the discussion of this topic would take us too far from the main theme of the paper we shall refer the reader to the literature [35,37].

Contrary to the deterministic situation minimal Markovian splitting subspaces are *nonunique*. Two very important examples are the *forward and backward predictor spaces* (at time zero):

$$\mathbf{X}_{-} := \mathrm{E}^{\mathbf{H}^{-}}\mathbf{H}^{+} \quad \mathbf{X}_{+} := \mathrm{E}^{\mathbf{H}^{+}}\mathbf{H}^{-} \tag{34}$$

for which we have the following characterization [35].

Proposition 3 *The subspaces* $\mathbf{X}_{-}$ *and* $\mathbf{X}_{+}$ *are minimal Markovian splitting subspaces contained in the past* $\mathbf{H}^{-}$*, and, respectively, in the future* $\mathbf{H}^{+}$*, of the process y.*

A basis in the forward predictor space $\mathbf{X}_{-}$ originates a stationary state-space model in which the state variables are linear functionals of the past history of the process y, i. e. $x(t) \in \mathbf{H}_t^{-}(y)$. In other words the state coincides with its best estimate (the orthogonal projection), $E[x(t) \mid \mathbf{H}_t^{-}(y)]$ given the past of y. It follows that the dynamical equations (8a) describe in this case a steady-state Kalman predictor and the input white noise $w = w_{-}$ is the steady state *innovation process* of y.

Before closing this section we should remark that a scattering picture emerges also in some of the early papers on the "unitary dilation" approach to statistical mechanics, e. g. [31,32]. It is remarkable that the abstract scattering picture described in this section makes contact with this literature. The dilation approach, however, deals (in our language) with the reverse problem of describing the stationary shift group starting from (Markov) processes which are described by a Langevin-type equation. This is a curious and (in our view) somewhat unnatural point of view which apparently does not lead to capture the phenomenon of *nonuniqueness of the macroscopic description* of the observables. Whether this has any physical relevance is however not clear to us.

Dynamical System Identification

System identification is the problem of describing an observed time series (e. g. a sequence of real numbers representing stock prices, temperatures in a room, sampled audio or video signals, etc.) by a linear dynamic model, in particular by a state-space model of the type (8a). It is a widely accepted viewpoint that system identification should be regarded as a *statistical* problem [40,42]. This is so because the experimental data which one wants to model very often come from complicated, often unknown, nonlinear physical phenomena subject to unknown interactions with the environment, and it would in general be impossible to attempt a "physical" modelization from first principles. Consequently the models describing the data should be stochastic. The general goal is to get accurate models which are as simple as possible. The problem can then be approached from (at least) two conceptually different viewpoints.

Identification by Parametric Optimization

This is the mainstream "optimization" approach, based on the principle of minimizing a suitable measure of discrepancy between the observed data and the data predicted by a probability law underlying a certain chosen model class. Well-known examples of distance functions are the *likelihood function*, or the average squared *prediction-error* of the observed data corresponding to a particular model. Except for rather trivial model classes, these distance functions depend nonlinearly on the model parameters and the minimization can only be done numerically. Hence the optimization approach leads to iterative algorithms in the parameter space, say in the space of minimal (A, B, C, D) matrix quadruples which parametrize a chosen model class. In spite of the fact that this has been almost the only accepted paradigm in system identification for many decades, [42,61], this approach has several drawbacks, including the need of unique parametrization of multivariable models, the fact that the cost function generally has complicated local minima which, for moderate or large dimension of the model are very difficult to detect, and the inherent insensitivity of the cost to variations of some parameters and the corresponding ill-posedness of the estimation problem.

It seems that these limitations are a consequence of the intrinsically "blind" philosophy which underlies setting the problem as a parameter optimization problem. Almost all problems of controller and/or estimator design could in principle be formulated just as parametric optimization after choosing a certain family of parametric structures for the controller or the estimator. Pushing this philosophy to the extreme, in principle one would not need the maximum principle, Kalman filtering, H^∞ theory, etc. (in fact one would not need system and control theory altogether), since everything could be reduced to a nonlinear programming problem in a suitable space of controller or estimator parameters. It is dubious however, whether any real progress in the field of control and estimation could have occurred by following this type of paradigm.

Identification by Stochastic Realization

This is commonly referred to as "subspace methods" identification or "subspace identification" *tout-court*. Subspace identification is essentially a sample version of stochastic realization. Under a reasonable assumption of second-order ergodicity of the process y which has produced the observed data, one can set up a "sample" counterpart of the abstract Hilbert space operations of stochastic realization theory. In this setting the geometric operations of stochastic realization can be translated into statistical operations on the observed data. One obtains a statistical theory of model building which is in a sense perfectly isomorphic to the abstract probabilistic realization theory. The main point is the idea of constructing first a (sample) *state space* for the process y, starting from certain vector spaces, called the *future* and *past* spaces associated with the observations. According to the recipe of (34) the state space is constructed by orthogonal projection of the future onto the past, as seen in the previous section. Successively, a well conditioned basis is chosen in the state space e. g. by principal components (canonical correlation) analysis. Once a "robust" basis; i. e. a state vector, is chosen, the parameters $A, C, \bar{C}$ of the model are computed by formulas analogous to (31),(32),(33). The final step of finding the B and D matrices requires solving a Riccati equation.

By similarity with realization theory, one recognizes that the inherent nonlinearity of model identification has to do with the quadratic nature of the spectral factorization problem. As we have seen, spectral factorization for state-space models involves the solution of a Riccati equation (or more generally of a linear matrix inequality), a problem which has been the object of intensive theoretical and numerical studies in the past three decades. The nonlinearity of the stochastic system identification problem is hence of a well-known and well-understood kind and is much better dealt with by the explicit methods of Riccati solution developed in system theory rather than by nonspecific optimization algorithms. This has allowed a systematic introduction of very reliable and efficient numerical linear algebra tools to solve the problem. The "subspace" approach has been suggested more or less implicitly by several authors in the past [6,13,15,44] but is first clearly presented in [63]. Various extensions to cover identification of systems with inputs have appeared since this paper and there is now a large literature on the subject. For references we refer the reader to the recently published book [27].

Since the optimization approach is well covered in the literature, we shall just analyze subspace identification below.

The Hilbert Space of a Time Series

The crucial conceptual step in subspace identification is a proper identification of the Hilbert space of random data in which modeling takes place. We shall briefly illustrate this point, following [38].

We shall initially consider an idealized situation in which the observed data (time series)

$$\{y_0, y_1, \ldots, y_t, \ldots\} \qquad y_t \in \mathbb{R}^m \tag{35}$$

is infinitely long.

We shall assume that the limit for $N \to \infty$ and for any $\tau \geq 0$, of the time averages

$$\frac{1}{N+1} \sum_{t=0}^{N} y_{t+\tau} y_t' \tag{36}$$

exists. It can be shown that this limit is a function $\tau \to \Lambda_\tau$ of *positive type*. In the continuous-time setting, functions admitting an "ergodic" limit of the sample correlation function (36), have been studied in depth by Wiener in his famous work on Generalized Harmonic Analysis. Although a systematic translation of the continuous-time results of Wiener into discrete-time seems not to be available in the literature, it is quite evident that a totally analogous set of results holds also for discrete-time signals. In particular it is rather easy to show, by adapting Wiener's proof for continuous time, that the limits of the time averages (36) form a matrix function Λ of positive type, in other words a *bona-fide stationary covariance matrix*, see [67,68] which can then be identified with the "true" covariance function of an underlying stochastic process.

The assumption actually implies that

$$\frac{1}{N+1} \sum_{t=t_0}^{N+t_0} y_{t+\tau} y_t' \to \Lambda_\tau \tag{37}$$

for arbitrary t_0, which can be read as a kind of "statistical regularity" of the (future) data. The ergodicity (36) is of course unverifiable in practice as it says something about data which have not been observed yet. Some assumption of this sort about the mechanism generating future data seems however to be necessary to even formulate the identification problem. We shall call Λ the *true covariance* of the time series $\{y_t\}$.

Now, for each $t \in \mathbb{Z}_+$ define the $m \times \infty$ matrices

$$\mathbb{y}(t) := [y_t, y_{t+1}, y_{t+2}, \ldots] \tag{38}$$

and consider the sequence $\mathbb{y} := \{\mathbb{y}(t) \mid t \in \mathbb{Z}_+\}$. We shall make this sequence of semi-infinite matrices into an object isomorphic to the stationary processes y.

Define the vector space $\mathbb{Y}$ of scalar semi-infinite real sequences obtained as finite linear combinations of the rows of y,

$$\mathbb{Y} := \left\{ \sum a_k' \mathbb{y}(t_k); \quad a_k \in \mathbb{R}^m,\ t_k \in \mathbb{Z}_+ \right\} \tag{39}$$

This space can be naturally made into an inner product space in the following way.

First, define the bilinear form $\langle \cdot, \cdot \rangle$ on the generators by letting

$$\langle a'\mathbb{y}(k), b'\mathbb{y}(j) \rangle := \lim_{N\to\infty} \frac{1}{N+1} \sum_{t=0}^{N} a' y_{t+k} y_{t+j}' b = a' \Lambda_{k-j} b, \tag{40}$$

for $a, b \in \mathbb{R}^m$, and then extend it by linearity to all elements of $\mathbb{Y}$.

Let $\mathbf{a} := \{a_k, k \in \mathbb{Z}_+\}$ be a sequence of vectors $a_k \in \mathbb{R}^m$, with compact support in $\mathbb{Z}_+$, and let $\mathbb{a}' := \{a_k'\}$. A generic element ξ of the vector space $\mathbb{Y}$ can be represented as

$$\xi = \sum_k a_k' \mathbb{y}(k) = \mathbf{a}' \mathbb{y}$$

Let us assume that the infinite block-symmetric Toeplitz matrix

$$T = \begin{bmatrix} \Lambda_0 & \Lambda_1 & \ldots & \Lambda_k & \ldots \\ \Lambda_1' & \Lambda_0 & \Lambda_1 & \ldots & \ldots \\ \vdots & & \ddots & & \vdots \\ \Lambda_k' & & & \Lambda_0 & \\ \ldots & & & & \end{bmatrix} \tag{41}$$

constructed from the "true" covariance sequence $\{\Lambda_0, \Lambda_1, \ldots, \Lambda_k, \ldots\}$ of the data, is positive definite. Since the bilinear form (40) on $\mathbb{Y}$ can be represented by the quadratic form

$$\langle \xi, \eta \rangle = \langle \mathbf{a}'\mathbb{y}, \mathbf{b}'\mathbb{y} \rangle = \sum_{kj} a_k' \Lambda_{k-j} b_j = \mathbf{a}' T \mathbf{b}$$

it can be seen that the bilinear form is nondegenerate (unless $\Lambda = 0$ identically) and defines a bona-fide inner product. As usual one identifies elements whose difference has norm zero (this means $\langle \xi, \xi \rangle = 0 \Leftrightarrow \xi = 0$).

By closing the vector space $\mathbb{Y}$ with respect to the norm induced by the inner product (40), one obtains a real Hilbert space of semi-infinite sequences which hereafter we shall still denote $\mathbb{Y}$. This is the basic data space on which our models will be defined.

The *shift operator* U operates on the family of semi-infinite matrices (38), by the rule

$$Ua'\mathbb{y}(t) = a'\mathbb{y}(t+1) \qquad t \in \mathbb{Z}\,, \quad a \in \mathbb{R}^m$$

It is easy to see that U is a linear map which is isometric with respect to the inner product (40) and can be extended by linearity to all of $\mathbb{Y}$. This Hilbert space framework for time series was introduced in [38]. It is shown in this reference that the "stationary Hilbert space" $(\mathbb{Y}, U)$ is isomorphic to the standard stochastic Hilbert space setup in the L^2 theory of second-order stationary random processes. By virtue of this isomorphism one can formally think of the observed time series (35) as an ergodic sample path of some Gaussian stationary stochastic process $\mathbf{y}$ defined on a true probability space, having covariance matrices equal to the limit Λ of the sum (36) as $N \to \infty$.

Linear functions and operators on the "tail sequences" $\mathbb{y}(t)$ correspond to the same linear functions and operators on the random variables of the process y. In particular the second-order moments of y can be computed in terms of the tail sequences $\mathbb{y}$, by substituting expectations with ergodic limits of the type (40). Since second-order properties are all what matters in our setting, one may even regard the tail sequence $\mathbb{y}$ of (38) as being the *same object* as the underlying stochastic process y. The usual probabilistic language can be adopted in the present setting provided one identifies real random variables as semi-infinite strings of numbers having the "ergodic property" described at the beginning of this section. The inner product of two semi-infinite strings ξ and η in $\mathbb{Y}$ corresponds in particular to the expectation $\mathbb{E}\{\xi\eta\}$; i. e.

$$\langle \xi, \eta \rangle = \mathbb{E}\{\xi\eta\}. \tag{42}$$

This unification of language permits us to carry over in its entirety the geometric theory of stochastic realization derived in the abstract L^2 setting to the time series framework. In particular, the *past* and *future* subspaces of the "processes" $\mathbb{y}$ at time t are defined as the closure of the linear vector spaces spanned by the relative past or future "random variables" $\mathbb{y}(t)$, in the metric of the Hilbert space $\mathbb{Y}$. The only difference to keep in mind here is the different interpretation that representation formulas like (8a) have in this context. The equalities involved in the representation

$$\begin{cases} x(t+1) = Ax(t) + Bw(t) \\ y(t) = Cx(t) + Dw(t) \end{cases} \tag{43}$$

are now to be understood in the sense of equalities of elements of $\mathbb{Y}$, i. e. as asymptotic equality of sequences in the sense of Cesàro limits. In particular the equality signs in the model (43) do not necessarily imply that the same relations hold for the sample values y_t, x_t, w_t at a particular instant of time t. This is in a certain sense similar to the "with probability one" interpretation of the equality sign to be given to the model (43) in case the variables are bona-fide random variables in a probability space.

Consider the orthogonal projection $\mathbb{E}[\,\xi \mid \mathbf{X}]$ of a (row) random variable ξ onto a subspace $\mathbf{X}$ of the space $\mathbb{Y}$. In the probabilistic L^2 setting this has the well-known interpretation of wide-sense conditional expectation given the random variables in $\mathbf{X}$ (a true conditional expectation, in the case of Gaussian distributions). In this setting the projection operator has an immediate and useful *statistical* meaning.

Assume for simplicity that $\mathbf{X}$ is given as the rowspace of some matrix of generators X, then the projection $\mathbb{E}[\,\xi \mid \mathbf{X}]$ has exactly the familiar aspect of the least squares formula expressing the best approximation of the vector ξ as a linear combination of the rows of X. For, writing $\mathbb{E}[\,\xi \mid X\,]$ to denote the projection expressed (perhaps nonuniquely) in terms of the rows of X, the classical linear "conditional expectation" formula leads to

$$\mathbb{E}[\,\xi \mid X\,] = \xi X'[XX']^{\sharp} X\,, \tag{44}$$

which is the universally known "least squares" formula of statistics. The pseudoinverse $\sharp$ can be substituted by a true inverse in case the rows of X are linearly independent.

The Question of Statistical Efficiency

There are questions about the statistical significance (what are the uncertainty bounds on the parameters and on the estimated transfer functions etc.) which are easily and naturally addressed in the optimization framework but harder to be addressed in the subspace/realization approach to identification. The main difference with the mainstream statistical approach is that the estimation of $(A, C, \bar{C})$ is not done *directly* by optimizing a likelihood or other distance functions but by just matching second-order moments, i. e. by solving the Eqs. (46).

This way of proceeding can be seen as an instance of *estimation by the method of moments* described in the statistical textbooks e. g. p. 497 in [12], a very old idea used extensively by K. Pearson in the beginning of the last century. The underlying estimation principle is that the parameter estimates should match exactly the sample second-order moments and is close in spirit to the wide-sense setting that we are working in. It does not involve optimality or minimal distance criteria between the "true" and the model distributions. For this reason it is generally claimed in the literature that one should expect better results (in the sense of smaller asymptotic variance of the estimates) by optimization methods. However, a proof that subspace methods can be efficient (under certain conditions which will be too long to report here) has appeared recently [8].

Usable expressions for the asymptotic covariance of the parameter estimates are given in [9]. The arguments in the derivations are, however, far from trivial.

Subspace Algorithms

Most standard algorithms for subspace identification can be shown to be essentially equivalent to the following three-step procedure (see e. g. [6,38]),

1. The first step is the estimation of a *finite* sequence of sample covariance matrices

$$\{\hat{\Lambda}_0, \hat{\Lambda}_1, \dots, \hat{\Lambda}_\nu\} \tag{45}$$

   from the observed data. Since the data are necessarily finite, ν must be also finite (and in general small compared to the data length).
2. The second step is identification of a rational model for the covariance sequence. This is a *minimal partial realization* (also called "rational extension") problem. Given a finite set of experimental covariance data one is asked to find a minimal value of n and a *minimal* triplet of matrices $(A, C, \bar{C})$, of dimensions $n \times n$, $m \times n$ and $m \times n$ respectively, such that

$$\hat{\Lambda}_k = CA^{k-1}\bar{C}' \qquad k = 1, \dots, \nu. \tag{46}$$

   Recall that $(A, C, \bar{C}')$ is a *minimal triplet*, in the sense of deterministic linear system theory, if the pairs of matrices (A, C) and $(A, \bar{C}')$ are an observable and, respectively, a reachable pair, see e. g. [24]. There are well-known algorithms for computing minimal partial realizations (see e. g. [72] for an efficient algorithm) thereby producing "estimates" of the parameters $(A, C, \bar{C})$ of a minimal realization of a rational spectral density matrix of the process.
3. The third step is to compute, starting from the realization of the rational spectrum estimated in step two, a stationary state-space model. Typically one is interested in the innovation model. This is accomplished by computing the minimal solution P_- of the Linear Matrix Inequality (16), or, equivalently, the minimal solution of the associated algebraic Riccati equation.

A warning is in order concerning the practical use of these algorithms in that some nontrivial mathematical questions related to positivity of the estimated spectrum are often overlooked in the implementation. This issue is thoroughly discussed in [38] and here we shall just give a short summary.

In determining a minimal triplet $(A, C, \bar{C})$ interpolating the partial sequence (45) so that $CA^{k-1}\bar{C}' = \hat{\Lambda}_k \quad k = 1, 2, \dots, \nu$, we also completely determine the infinite sequence

$$\{\hat{\Lambda}_0, \hat{\Lambda}_1, \hat{\Lambda}_2, \hat{\Lambda}_3, \dots\} \tag{47}$$

by setting $\hat{\Lambda}_k = CA^{k-1}\bar{C}'$ for $k = \nu+1, \nu+2, \dots$. This sequence is called a *minimal rational extension* of the finite sequence (45). The attribute "rational" is due to the fact that

$$\begin{aligned}\hat{\Phi}_+(z) &:= \frac{1}{2}\hat{\Lambda}_0 + \hat{\Lambda}_1 z^{-1} + \hat{\Lambda}_2 z^{-2} + \dots \\ &= \frac{1}{2}\hat{\Lambda}_0 + C(zI - A)^{-1}\bar{C}'\end{aligned} \tag{48}$$

is a rational function. In order for (47) to be a bona fide covariance sequence, however, it is necessary that the *infinite* block-Toeplitz matrix obtained by extending the finite matrix

$$T_\nu = \begin{bmatrix} \hat{\Lambda}_0 & \hat{\Lambda}_1 & \hat{\Lambda}_2 & \cdots & \hat{\Lambda}_\nu \\ \hat{\Lambda}_1' & \hat{\Lambda}_0 & \hat{\Lambda}_1 & \cdots & \hat{\Lambda}_{\nu_1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{\Lambda}_\nu' & \hat{\Lambda}_{\nu-1}' & \hat{\Lambda}_{\nu-2} & \dots & \hat{\Lambda}_0 \end{bmatrix} \tag{49}$$

using the rational extension sequence (47), should be nonnegative definite. Equivalently, the spectral density function corresponding to (47)

$$\hat{\Phi}(z) = \hat{\Lambda}_0 + \sum_{k=1}^{\infty} \hat{\Lambda}_k (z^k + z^{-k}) = \hat{\Phi}_+(z) + \hat{\Phi}_+(z^{-1})' \tag{50}$$

should be nonnegative definite on the unit circle. This is equivalent to the function $\hat{\Phi}_+(z)$ being *positive real*. Consequently, the partial realization needs to be done subject to the extra constraint of positivity.

The constraint of positivity is a rather tricky one and in some of the methods described in the literature it is disregarded. For this reason some subspace algorithms may fail to provide a positive extension and hence may lead to data $(A, C, \bar{C})$ for which there are no solutions of the LMI and hence to totally inconsistent results.

It is important to appreciate the fact that the problem of positivity of the extension has little to do with the noise or sample variability of the covariance data and is present equally well for any *finite* covariance sequence extracted from a true (infinitely long) rational covariance sequence. For there is no guarantee that, even in this idealized situation, the order of a minimal rational extension (47) of the finite covariance subsequence would be sufficiently large to equal the order of the infinite sequence and hence to generate a positive extension. A minimal partial realization may well fail to be positive because its order is too low to guarantee positivity. It is shown in [38] that there are relatively "large" sets of data (45) for which this happens and the matrix A may even fail to be stable [7].

Future Directions

Because of the pervasive presence of fast digital computers in modern society, we are witnessing a sharp change of paradigms in the analysis, design, prediction and control of both man made and natural (say biological, economical etc) systems. Because of the tremendous computing capabilities at our disposal nowadays, stochastic modeling, realization and identification have become an essential ingredient of modern applied sciences. In particular modeling and simulation of complex technological, biological, economic, and environmental systems is becoming more and more essential for understanding the behavior and for prediction and control of these systems. New problems continually arise and much work remains to be done.

We just mention here problems where there is a need of modeling the effect of *input* or *decision variables*, possibly in the presence of feedback. Such *Input–Output* models have important applications in many applied areas such as control of industrial processes, econometrics, etc. There is a large body of results concerning stochastic realization and subspace identification of systems with inputs which we could not discuss in this article, see [10,11,26,27,53,54,55,64,65]. The area is important for applications and still needs research. Modeling and realization of finite state processes and of certain classes of highly non-Gaussian signals is needed in diverse applications such as source coding, image processing, and computer vision.

Bibliography

Primary Literature

1. Akaike H (1975) Markovian representation of stochastic processes by canonical variables. SIAM J Control 13:162–173
2. Akaike H (1974) Stochastic Theory of Minimal Realization. IEEE Trans Autom Control AC-19(6):667–674
3. Anderson BDO (1969) The inverse problem of Stationary Covariance Generation. J Stat Phys 1(1):133–147
4. Anderson BDO (1967) An algebraic solution to the spectral factorization problem. IEEE Trans Autom Control AC-12:410–414
5. Anderson BDO (1999) The realization problem for Hidden Markov Models. Math Control Signals Syst 12:80–120
6. Aoki M (1990) State Space Modeling of Time Series, 2nd edn. Springer, New York
7. Byrnes CI, Lindquist A (1982) The stability and instability of partial realizations. Syst Control Lett 2:2301–2312
8. Bauer D (2005) Comparing the CCA Subspace Method to Pseudo Maximum Likelihood Methods in the case of No Exogenous Inputs. J Time Ser Anal 26:631–668
9. Chiuso A, Picci G (2004) The Asymptotic Variance of Subspace Estimates. J Econom 118(1–2):257–291
10. Chiuso A, Picci G (2004) On the Ill-conditioning of subspace identification with inputs. Automatica 40(4):575–589
11. Chiuso A, Picci G (2005) Consistency Analysis of some Closed-loop Subspace Identification Methods. Automatica 41:377–391, special issue on System Identification
12. Cramèr H (1949) Mathematical Methods of Statistics. Princeton University Press, Princeton
13. Desai UB, Pal D, Kirkpatrick RD (1985) A Realization Approach to Stochastic Model Reduction. Int J Control 42:821–838
14. Doob JL (1953) Stochastic Processes. Wiley, New York
15. Faurre P (1969) Identification par minimisation d'une representation Markovienne de processus aleatoires. Symposium on Optimization, Nice 1969. Lecture Notes in Mathematics, vol 132. Springer, New York
16. Faurre P, Chataigner P (1971) Identification en temp reel et en temp differee par factorisation de matrices de Hankel. French–Swedish colloquium on process control, IRIA Roquencourt
17. Faurre P (1973) Realisations Markovienne de processus stationnaires. Report de Recherche no 13, INRIA, Roquencourt
18. Ferrante A, Picci G, Pinzoni S (2002) Silverman algorithm and the structure of discrete-time stochastic systems. Linear Algebra Appl (special issue on systems and control) 351–352:219–242
19. Finesso L, Spreij PJC (2002) Approximate realization of finite Hidden Markov Chains. Proceedings of the 2002 IEEE Information Theory Workshop, Bangalore, pp 90–93
20. Ford GW, Kac M, Mazur P (1965) Statistical mechanics of assemblies of coupled oscillators. J Math Phys 6:504–515
21. Kalman RE (1963) On a new Criterion of Linear Passive Systems. In: Proc. of the First Allerton Conference. University of Illinois, Urbana Ill, Nov 1963, pp 456–470
22. Kalman RE (1963) Lyapunov Functions for the problem of Lur'e in Automatic Control. Proc Natl Acad Sci 49:201–205
23. Kalman RE (1965) Linear stochastic filtering: reappraisal and outlook. In: Proc. Symposium on Syatem Theory, Politechnic Institute of Brooklin, pp 197–205
24. Kalman RE, Falb PL, Arbib MA (1969) Topics in Mathematical Systems Theory. McGraw–Hill, New York

25. Karatzas I, Shreve S (1987) Brownian motion and stochastic calculus. Springer, New York
26. Katayama T, Picci G (1999) Realization of Stochastic Systems with Exogenous Inputs and Subspace Identification Methods. Automatica 35(10):1635–1652
27. Katayama T (2005) Subspace methods for System Identification. Springer, New York
28. Kintchine A (1934) Korrelationstheorie Stationäre Stochastischen Prozessen. Math Annalen 109:604–615
29. Lax PD, Phillips RS (1967) Scattering Theory. Academic Press, New York
30. Levy P (1948) Processus stochastiques et Mouvement Brownien. Gauthier–Villars, Paris
31. Lewis JT, Thomas LC (1975) How to make a heat bath. In: Arthurs AM (ed) Functional Integration and its Applications. Clarendon, Oxford
32. Lewis JT, Maassen H (1984) Hamiltonian models of classical and quantum stochastic processes. In: Accardi L, Frigerio A, Gorini V (eds) Quantum Probability and Applications to the Quantum Theory of Irreversible processes. Springer Lecture Notes in Mathematics, vol 1055. Springer, New York, pp 245–276
33. Lindquist A, Picci G, Ruckebusch G (1979) On minimal splitting subspaces and Markovian representation. Math Syst Theory 12:271–279
34. Lindquist A, Picci G (1979) On the stochastic realization problem. SIAM J Control Optim 17:365–389
35. Lindquist A, Picci G (1985) Realization theory for multivariate stationary Gaussian processes. SIAM J Control Optim 23:809–857
36. Lindquist A,Pavon M (1984) On the structure of state space models of discrete-time vector processes. IEEE Trans Autom Control AC-29:418–432
37. Lindquist A, Picci G (1991) A geometric approach to modelling and estimation of linear stochastic systems. J Math Syst Estim Control 1:241–333
38. Lindquist A, Picci G (1996) Canonical correlation analysis approximate covariance extension and identification of stationary time series. Automatica 32:709–733
39. Lindquist A, Michaletzky G, Picci G (1995) Zeros of Spectral Factors, the Geometry of Splitting Subspaces, and the Algebraic Riccati Inequality. SIAM J Control Optim 33:365–401
40. Lindquist A, Picci G (1996) Geometric Methods for State Space Identification. In: Identification, Adaptation, Learning. (Lectures given at the NATO-ASI School, From Identification to Learning, Como, Italy, Aug 1994. Springer
41. Lindquist A, Picci G, Linear Stochastic Systems: A Geometric Approach. (in preparation)
42. Ljung L (1987) System Identification – Theory for the User. Prentice–Hall, New York
43. Martin-Löf A (1979) Statistical Mechanics and the foundations of Thermodynamics. Lecture Notes in Physics, vol 101. Springer, New York
44. Moonen M, De Moor B, Vanderberghe L, Vandewalle J (1989) On- and Off-Line Identification of Linear State–Space Models. Int J Control 49:219–232
45. Taylor TJ, Pavon M (1988) A solution of the nonlinear stochastic realization problem. Syst Control Lett 11:117–121
46. Picci G (1978) On the internal structure of finite-state stochastic processes. In: Mohler R, Ruberti A (eds) Recent developments in Variable Structure Systems. Springer Lecture Notes in Economics and Mathematical Systems. vol 162. Springer, New York, pp 288–304
47. Picci G, Pinzoni S (1994) Acausal Models and Balanced realizations of stationary processes. Linear Algebra Appl (special issue on Systems Theory) 205–206:957–1003
48. Picci G (1986) Application of stochastic realization theory to a fundamental problem of statistical physics. In: Byrnes CI, Lindquist A (eds) Modeling, Identification and Robust Control. North Holland, Amsterdam
49. Picci G, Taylor TJ (1990) Stochastic aggregation of linear Hamiltonian systems with microcanonical distribution. In: Kaashoek MA, van Schuppen JH, Ran ACM (eds) Realization and modeling in System Theory. Birkhäuser, Boston, pp 513–520
50. Picci G (1992) Markovian representation of linear Hamiltonian systems. In: Guerra F, Loffredo MI, Marchioro C (eds) Probabilistic Methods in Mathematical Physics. World Sciedntific, Singapore, pp 358–373
51. Picci G (1991) Stochastic realization theory. In: Antoulas A (ed) Mathematical System Theory. Springer, New York
52. Picci G, Taylor TSJ (1992) Generation of Gaussian Processes and Linear Chaos. In: Proc 31st IEEE Conf on Decision and Control. Tucson, pp 2125–2131
53. Picci G, Katayama T (1996) Stochastic realization with exogenous inputs and "Subspace Methods" Identification. Signal Process 52(2):145–160
54. Picci G (1996) Geometric methods in Stochastic Realization and System Identification. CWI Q (invited paper) 9:205–240
55. Picci G (1997) Oblique Splitting Susbspaces and Stochastic Realization with Inputs. In: Helmke U, Prätzel–Wolters D, Zerz E (eds) Operators, Systems and Linear Algebra. Teubner, Stuttgart, pp 157–174
56. Popov VM (1964) Hyperstability and Optimality of Automatic Systems with several Control functions. Rev Rumaine Sci Tech Ser Electrotech Energ 9:629–690
57. Rozanov NI (1963) Stationary Random Processes. Holden-Day, San Francisco
58. Ruckebusch G (1976) Répresentations markoviennes de processus gaussiens stationnaires. Comptes Rendues Acad Sci Paris Series A 282:649–651
59. Ruckebusch G (1978) A state space approach to the stochastic realization problem. Proc 1978 IEEE Intern Symp Circuits and Systems, pp 972–977
60. Ruckebusch G (1978) Factorisations minimales de densités spectrales et répresentations markoviennes. In: Proc 1re Colloque AFCET–SMF, Palaiseau
61. Söderström T, Stoica P (1989) System Identification. Prentice–Hall, New York
62. van Schuppen JH (1989) Stochastic realization problems. In: Nijmeijer H, Schumacher JM (eds) Three decades of mathematical system theory. Lecture Notes in Control and Information Sciences, vol 135. Springer, New York, pp 480–532
63. Van Overschee P, De Moor B (1993) Subspace algorithms for the stochastic identification problem. Automatica 29:649–660
64. Van Overschee P, De Moor B (1994) N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems. Automatica 30:75–93
65. Verhaegen M (1994) Identification of the deterministic part of MIMO State Space Models given in Innovations form from Input–Output data. Automatica 30:61–74
66. Vidyasagar M, (Hidden) Markov Models: Theory and Applications to Biology. Princeton University Press, Princeton

67. Wiener N (1930) Generalized Harmonic Analysis. Acta Math 55:117–258
68. Wiener N (1933) Generalized Harmonic Analysis. In: The Fourier Integral and Certain of its Applications. Cambridge University Press, Cambridge
69. Wiener N (1949) Extrapolation, Interpolation and Smoothing of Stationary Time Series. The M.I.T. Press, Cambridge
70. Willems JC (1971) Least Squares Stationary Optimal Control and the Algebraic Riccati Equation. IEEE Trans Autom Control AC-16:621–634
71. Yakubovich VA (1963) Absolute stability of Nonlinear Control Systems in critical case: part I and II. Avtom Telemeh 24:293–303, 717–731
72. Zeiger HP, McEwen AJ (1974) Approximate linear realization of given dimension via Ho's algorithm. IEEE Trans Autom Control AC-19:153

Books and Reviews

Doob JL (1942) The Brownian movement and stochastic equations. Ann Math 43:351–369

Faurre P, Clerget M, Germain F (1979) Opérateurs Rationnels Positifs. Dunod, Paris

Kalman RE (1981) Realization of covariance sequences. Proc Toeplitz Memorial Conference, Tel Aviv, Birkhäuser

Nelson E (1967) Dynamical theories of Brownian motion. Princeton University press, Princeton

Picci G (1976) Stochastic realization of Gaussian processes. Proc IEEE 64:112–122

Picci G (1992) Stochastic model reduction by aggregation. In: Isidori A, Tarn TJ (eds) Systems Models and Feedback: Theory and Applicatons. Birkhäuser Verlag, Basel, pp 169–177

Stochastic Processes

Alan J. McKane
Theory Group, School of Physics and Astronomy,
University of Manchester, Manchester, UK

Article Outline

Glossary

Fokker–Planck equation A partial differential equation of the second order for the time evolution of the probability density function of a stochastic process. It resembles a diffusion equation, but has an extra term which represents the deterministic aspects of the process.

Langevin equation A stochastic differential equation of the simplest kind: linear and with an additive Gaussian white noise. Introduced by Langevin [30] in 1908 to describe Brownian motion; many stochastic differential equations in physics go by this name.

Markov process A stochastic process in which the current state of the system is only determined from its state in the immediate past, and not by its entire history.

Markov chain A Markov process where both the states and the time are discrete and where the process is stationary.

Master equation The equation describing a continuous-time Markov chain.

Stochastic process A sequence of stochastic variables. This sequence is usually a time-sequence, but could also be spatial.

Stochastic variable A random variable. This is a function which maps outcomes to numbers (real or integer).

Definition of the Subject

The most common type of stochastic process comprises of a set of random variables $\{x(t)\}$, where t represents the time which may be real or integer valued. Other types of stochastic process are possible, for instance when the stochastic variable depends on the spatial position $\mathbf{r}$, as well as, or instead of, t. Since in the study of complex systems we will predominantly be interested in applications relating to stochastic dynamics, we will suppose that it depends only on t. One of the earliest investigations of a stochastic process was carried out by Bachelier [2], who used the idea of a random walk to analyze stock market fluctuations. The problem of a random walk was more generally discussed by Pearson [42], and applied to the investigation of Brownian motion by Einstein [8,9], Smoluchowski [54] and Langevin [30]. The example of a random walk illustrates the fact that in addition to time being discrete or continuous, the stochastic variable itself can be discrete (for instance, the walker moves with fixed step size in one dimension) or continuous (for instance, the velocity of a Brownian particle). The modeling of the process may lead to an equation for the stochastic variable, such as a stochastic differential equation, or for an equation which predicts how the probability density function (pdf) for the stochastic variable changes in time. Stochastic processes are ubiquitous in the physical, biological and social sciences; they may come about through the perception of very complicated processes being essentially random (the toss of a coin, roll of a die, birth or death of individuals in populations), the inclusion of diverse and poorly characterized effects external to the system under consideration, or thermal fluctuations, among others.

Introduction

Deterministic dynamical processes are typically formulated as a set of rules which allow for the state of the system at time $t+1$ (or $t+\delta t$) to be found from the state of the system at time t. By contrast, for stochastic systems, we can only specify the probability of finding the system in a given state. If this only depends on the state of the system at the previous time step, but not those before this, the stochastic process is said to be Markov. Fortunately many stochastic processes are Markovian to a very good approximation, since the theory of non-Markov processes is considerably more complicated than Markov processes and much less well developed. In this article we will deal almost exclusively with Markov processes.

The mathematical definition of a Markov process follows from the definition of the hierarchy of pdfs for a given process. This involves the joint pdfs $P(x_1, t_1; x_2, t_2; \ldots; x_n, t_n)$, which are the probability that the system is in state x_1 at time t_1, state x_2 at time t_2,..., and state x_n at time t_n, and also the conditional pdfs $P(x_1, t_1; \ldots; x_m, t_m | x_{m+1}, t_{m+1}; \ldots; x_n, t_n)$, which are the probability that the system is in state x_1 at time $t_1, \ldots, x_m$ at time t_m, *given* that it was in state x_{m+1} at time $t_{m+1}, \ldots, x_n$ at time t_n. These pdfs are all non-negative and normalizable, and relations exist between them due to symmetry and reduction (integration over some of the state variables). Nevertheless, for a general non-Markov process, a whole family of these pdfs will be required to specify the process. On the other hand, for a Markov process the history of the system, apart from the immediate past, is forgotten, and so $P(x_1, t_1; \ldots; x_m, t_m | x_{m+1}, t_{m+1}; \ldots; x_n, t_n) = P(x_1, t_1; \ldots, x_m; t_m | x_{m+1}, t_{m+1})$. A direct consequence of this is that the whole hierarchy of pdfs can be determined from only two of them: $P(x,t)$ and $P(x,t|x',t')$. The hierarchy of defining equations then collapses to only two:

$$P(x_2, t_2) = \int \mathrm{d}x_1 P(x_2, t_2 | x_1, t_1) P(x_1, t_1) \quad (1)$$

and

$$P(x_3, t_3 | x_1, t_1) = \int \mathrm{d}x_2 P(x_3, t_3 | x_2, t_2) P(x_2, t_2 | x_1, t_1)\,, \quad t_1 < t_2 < t_3\,. \quad (2)$$

The pdf $P(x,t|x',t')$ is referred to as the transition probability and Eq. (2) as the Chapman–Kolmogorov equation. While the pdfs for a Markov process must obey Eqs. (1) and (2), the converse also holds: any two non-negative functions $P(x,t)$ and $P(x,t|x',t')$ which satisfy Eqs. (1) and (2), uniquely define a Markov process.

We will begin our discussion in Sect. "Markov Chains" with what is probably the simplest class of Markov processes: the case when both the state space and time are discrete. These are called Markov chains and were first investigated, for a finite number of states, by Markov [32] in 1906. The extension to an infinite number of states was carried out by Kolmogorov [27] in 1936. If time is continuous, an analogous formalism may be developed, which will be discussed in Sect. "The Master Equation". In physics the equation describing the time evolution of the pdf in this case is called the master equation and was introduced by Pauli [41] in 1928, in connection with the approach to equilibrium for quantum systems, and also by Nordsieck, Lamb and Uhlenbeck [39] in 1940, in connection with fluctuations in cosmic ray physics. The term "master equation" refers to that fact that many of the quantities of interest can be derived from this equation. The connection with previous work on Markov processes was clarified by Siegert [49] in 1949.

In many instances when the master equation cannot be solved exactly, it is useful to approximate it by a rather coarser description of the system, known as the Fokker–Planck equation. This approach will be discussed in Sect. "The Fokker–Planck Equation". This equation was used in its linear form by Rayleigh [44], Einstein [8,9], Smoluchowski [54,55], and Fokker [14], but it was Planck [43] who derived the general nonlinear form from a master equation in 1917, and Kolmogorov who made the procedure rigorous in 1931. All the descriptions which we have mentioned so far have been based on the time evolution of the pdfs. An alternative specification is to give the time evolution of the stochastic variables themselves. This will necessarily involve random variables appearing in the equations describing this evolution, and they will therefore be *stochastic differential equations*. The classic example is the Langevin equation [30] used to describe Brownian motion. This equation is linear and can therefore be solved exactly. The Langevin approach, and its relation to the Fokker–Planck equation is described in Sect. "Stochastic Differential Equations".

A good summary of the understanding of stochastic processes that had been gained by the mid-1950s is given in the book edited by Wax. This covers the basics of the subject, and what is discussed in the first six sections of this article. The article by Chandrasekhar [4], first published in 1943, and reprinted in Wax, gives an extensive bibliography of stochastic problems in physics before 1943. Since then the applications of the subject have grown enormously, and the equations modeling these systems have correspondingly become more complex. We illustrate some of the procedures which have been developed to deal with these equations in the next two sections. In Sect. "Path Integrals" we discuss how the path-integral formalism may be applied to stochastic processes and in Sect. "System Size Expansion" we describe how master equations can be analyzed when the size of the system is large. We end with a look forward to the future in Sect. "Future Directions".

Markov Chains

The simplest version of the Markov process is when both the states and the time are discrete, and when the stochastic process is *stationary*. When the states are discrete we will denote them by n or m, rather than x, which we reserve for continuous state variables. In this notation the two Eqs. (1) and (2) governing Markov processes read

$$P(n_2, t_2) = \sum_{n_1} P(n_2, t_2 | n_1, t_1) P(n_1, t_1) \tag{3}$$

$$P(n_3, t_3 | n_1, t_1) = \sum_{n_2} P(n_3, t_3 | n_2, t_2) P(n_2, t_2 | n_1, t_1) \,, \quad t_1 < t_2 < t_3 \,. \tag{4}$$

A stationary process is one in which the conditional pdf $P(n, t|n', t')$ only depends on the time difference $(t - t')$. For such processes, when time is discrete so that $t = t' + 1, t' + 2, \ldots$, we may write $P(n, t' + k|n', t')$ as $p^{(k)}_{n\,n'}$. The most elementary form of the Chapman–Kolmogorov equation (4) may then be expressed as

$$p^{(2)}_{n\,m} = \sum_{n'} p^{(1)}_{n\,n'} p^{(1)}_{n'\,m} \,. \tag{5}$$

This corresponds to the matrix multiplication of $p^{(1)}$ with itself, and therefore $p^{(2)}$ is simply $(p^{(1)})^2$. In the same way $p^{(k)} = (p^{(1)})^k$, and from now on we drop the superscript on $p^{(1)}$ and denote the matrix by $\mathcal{P}$. The entries of $\mathcal{P}$ are non-negative, with the sum of entries in each column being equal to unity, since

$$\sum_{n} p_{n\,n'} = \sum_{n} P(n, t+1|n', t) = 1 \,. \tag{6}$$

Such matrices are called *stochastic matrices*. From Eq. (5) it is clear that $\mathcal{P}^2$ is also a stochastic matrix, and by induction it follows that $\mathcal{P}^k$ is a stochastic matrix if $\mathcal{P}$ is.

The other defining relation for a Markov process, Eq. (3), now becomes

$$P(n, t+1) = \sum_{n'} p_{n\,n'} P(n', t) \,. \tag{7}$$

This relation defines a *Markov chain*. It has two ingredients: the probability that the system is in state n at time t, $P(n, t)$ – which is usually what we are trying to determine, and the stochastic matrix with entries $p_{n\,n'}$ which gives the probabilities of transitions from the state n' to the state n. The transition probabilities are typically given; they define the model. Note that in many texts the probability of making a transition from n' to n is written as $p_{n'\,n}$, not $p_{n\,n'}$. If we write $P(n, t)$ as a vector $\mathbf{P}(t)$, then we may write Eq. (7) as $\mathbf{P}(t+1) = \mathcal{P}\mathbf{P}(t)$. Therefore,

$$\mathbf{P}(t) = \mathcal{P}\mathbf{P}(t-1) = \mathcal{P}\mathcal{P}\mathbf{P}(t-2) = \cdots = \mathcal{P}^t\mathbf{P}(0) \,, \tag{8}$$

and so if the initial state of the system $\mathbf{P}(0)$ is given, then we can find the state of the system at time t ($\mathbf{P}(t)$) by matrix multiplication by the tth power of the transition matrix.

Examples of Markov Chains

1. *A one-dimensional random walk*. The most widely known example of a Markov chain is a random walk on the real axis, where the walker takes single steps between integers on the line. The simplest version is where the walker has to move during every time interval:

$$p_{n\,n'} = \begin{cases} p \,, & \text{if} \quad n = n' + 1 \\ q \,, & \text{if} \quad n = n' - 1 \\ 0 \,, & \text{otherwise} \,, \end{cases} \tag{9}$$

where $p + q = 1$. There are many variants. For instance, the walker could have a non-zero probability of staying put, in which case $p_{n\,n} = r$, with $p + q + r = 1$. The walk could be heterogeneous, in which case p and q (and r), could depend on n'. If there are boundaries, the boundary conditions have to be given. The most common two are *absorbing boundaries* defined by

$$\begin{aligned} p_{n+1\,n} &= p \,, \quad p_{n-1\,n} = q \,, \quad (n = 2, \ldots, N-1) \\ p_{11} &= 1 \,, \quad p_{NN} = 1 \,, \\ p_{n\,n'} &= 0 \,, \quad \text{otherwise} \,, \end{aligned} \tag{10}$$

and *reflecting boundaries* defined by

$$\begin{aligned} p_{n+1\,n} &= p\,, \quad p_{n-1\,n} = q\,, \quad (n = 2,\dots,N-1) \\ p_{21} &= p\,, \quad p_{N-1\,N} = q\,, \\ p_{11} &= q\,, \quad p_{NN} = p\,, \\ p_{n\,n'} &= 0\,, \quad \text{otherwise}\,. \end{aligned} \tag{11}$$

With absorbing boundaries (10), if we reach the state 1 or N, we can never leave it, since we stay there with probability 1. When the boundary is reflecting, we can never move beyond it; the only options are to move back towards the other boundary, or stay put. Well known examples of absorbing boundaries include the gambler's ruin problem, where a gambler bets a given amount against the house at each time step and can win with a probability p. Here the absorbing boundary is situated at $n = 0$. Eventually he will arrive at the state $n = 0$, where is has no money left, and so cannot re-enter the game. Birth/death processes will also have absorbing states at $n = 0$: if n is the number of individuals at a given time, and p is the probability of a birth and q of a death, then if there are no individuals left ($n = 0$), none can be born. This condition will be automatically applied if the transition probabilities are proportional to the number of individuals in the population.

2. *The Ehrenfest urn* This Markov chain was introduced by the Ehrenfests [7] in 1907, to illustrate the approach to equilibrium in a gas. Two containers, A and B, contain molecules of the same gas, the sum of the number of molecules in A and B being fixed to be N. At time t a molecule is removed at random from the containers and put into the other. If n' is the number of molecules in container A at a certain time, then at the next time step the transition probabilities will be:

$$p_{n\,n'} = \begin{cases} \frac{n'}{N}\,, & \text{if} \quad n' = n+1 \\ (1 - \frac{n'}{N})\,, & \text{if} \quad n' = n-1 \\ 0\,, & \text{otherwise}\,. \end{cases} \tag{12}$$

This is clearly a heterogeneous random walk of the type (9), and another interpretation of this model is as a random walk, but with a central force.

The most frequently asked question concerning Markov chains is: what is their eventual fate; how does the system behave at large time? Clearly if it tends towards a non-trivial stationary state, $P_{\text{st}}(n)$, then from Eq. (7):

$$P_{\text{st}}(n) = \sum_{n'} p_{n\,n'} P_{\text{st}}(n')\,, \tag{13}$$

and so $P_{\text{st}}(n)$ is a right eigenvector of $\mathcal{P}$ with unit eigenvalue. It follows from the properties of a general stochastic matrix that the eigenvalues of a stochastic matrix are such that $|\lambda| \le 1$ [15]. Furthermore every stochastic matrix has an eigenvalue equal to 1, however it may not be simple – there may be a multiplicity of unit eigenvalues. The classification of Markov chains can be used to decide which of these possibilities is the case. For example, Markov chains may be reducible or irreducible, and states recurrent or transient. We shall not discuss this in detail; Feller [10] gives a clear account of this classification and Cox and Miller [5] explore the consequences for the nature of the eigenvalues. Instead we will examine a specific example, that of the Ehrenfest urn introduced above, and focus on the explicit calculation of the eigenvalues and eigenvectors in that case.

Suppose that $\Psi^{(k)}$ and $\Theta^{(k)}$ are the right- and left-eigenvectors of $\mathcal{P}$, respectively, corresponding to the eigenvalue $\lambda^{(k)}$, so that $\Theta^{(k)}.\,\Psi^{(\ell)} = \delta_{k\,\ell}$. Then, in general, and for the Ehrenfest urn in particular,

$$\mathcal{P}^t_{n\,n'} = \sum_{k=0}^{N} \psi^{(k)}_n \left(\lambda^{(k)}\right)^t \theta^{(k)}_{n'}\,, \tag{14}$$

where $\psi^{(k)}_n$ is the nth component of the vector $\Psi^{(k)}$ and similarly for the left-eigenvector. The eigenvalues and eigenvectors for the Ehrenfest urn can be found exactly (Kac [23]; see also Krafft and Schaefer [28]). The eigenvalues are $\lambda^{(k)} = 1 - (2k/N)$, $k = 0,\dots,N$. Thus in this case there is a single eigenvalue $\lambda = 1$. The corresponding right-eigenvector, which gives the stationary state, is a binomial distribution:

$$P_{\text{st}}(n) = \psi^{(0)}_n = \frac{N!}{n!(N-n)!}\,\frac{1}{2^N}\,. \tag{15}$$

The left-eigenvector is $\theta^{(0)}_n = 1$ for all n. The other eigenvectors have the form $\theta^{(k)}_n = a_{kn}$ and $\psi^{(k)}_n = a_{kn} P_{\text{st}}(n)$, where the a_{kn} are the Krawtchouk polynomials [1]. Clearly, $a_{0n} = 1$, and the first non-trivial polynomial is $a_{1n} = \sqrt{N}[1 - (2n/N)]$. Therefore, for a suitable choice of initial conditions and using Eq. (8), the large t behavior of the Ehrenfest urn can be found from

$$\mathcal{P}^t_{n\,n'} \approx P_{\text{st}}(n) \left\{ 1 + c_{n\,n'} \left(1 - \frac{2}{N}\right)^t \right\}\,, \tag{16}$$

where $c_{n\,n'} = N[1 - (2n/N)][1 - (2n'/N)]$.

The Master Equation

The master equation is a Markov chain in the limit where time is continuous. To derive it we will assume that the states are discrete (the derivation is essentially identical if they are continuous) and write down the Chapman–Kolmogorov equation (4) in the form:

$$P(n,t+\Delta t|n_0,t_0) = \sum_{n'} P(n,t+\Delta t|n',t)P(n',t|n_0,t_0)\,. \tag{17}$$

We consider only stationary processes, so that we may take $t_0 = 0$ without loss of generality and $P(n, t + \Delta t|n', t)$ is independent of t. We now assume that

$$P(n,t+\Delta t|n',t) = \begin{cases} 1-\kappa(n)\Delta t + o(\Delta t)\,, & \text{if } n = n' \\ T(n|n')\Delta t + o(\Delta t)\,, & \text{if } n \neq n'\,, \end{cases} \tag{18}$$

where $o(\Delta t)$ means that $o(\Delta t)/\Delta t$ tends to zero as $\Delta t \to 0$. This reasonable: after a very short times the transition probability to stay put is unity minus a term of order Δt and the transition probabilities to move to any other state is of order Δt, but this is still an additional assumption on the process. The quantity $T(n|n')$ is the transition rate, and is only defined for $n \neq n'$. Since $\sum_n P(n, t + \Delta t|n', t) = 1$ for all n', we have that

$$\kappa(n') = \sum_{n\neq n'} T(n|n')\,. \tag{19}$$

Substituting Eq. (18) into Eq. (17), and making use of Eq. (19) we find that

$$\begin{aligned}&\frac{P(n,t+\Delta t|n_0,0) - P(n,t|n_0,0)}{\Delta t} \\ &= \sum_{n'\neq n} \left[T(n|n')P(n',t|n_0,0)\right] \\ &\quad - P(n,t|n_0,0)\sum_{n'\neq n}\left[T(n'|n)\right] + \frac{o(\Delta t)}{\Delta t}\,. \end{aligned} \tag{20}$$

Taking the limit $\Delta t \to 0$ gives *the master equation* for how the probability of finding the system in state n at time t changes with time:

$$\frac{\mathrm{d}P(n,t)}{\mathrm{d}t} = \sum_{n'\neq n} T(n|n')\,P(n',t) - \sum_{n'\neq n} T(n'|n)\,P(n,t)\,. \tag{21}$$

We have dropped the initial conditions, assuming that they are understood. It should be noticed that an analogous analysis starting from Eq. (3), rather than Eq. (4), may be carried out, leading to identical equations for $P(n, t|n_0, 0)$ and $P(n, t)$. If the state space is continuous the master equation reads

$$\frac{\partial P(x,t)}{\partial t} = \int \mathrm{d}x' \left[T(x|x')\,P(x',t) - T(x'|x)\,P(x,t)\right]\,. \tag{22}$$

In most applications transitions only take place between states whose label differs by one. That is, $T(n|n')$ and $T(n'|n)$ are zero unless $n' = n + 1$ and $n' = n - 1$. These are called *one-step processes*. For such processes the master equation takes the simpler form

$$\begin{aligned}\frac{\mathrm{d}P(n,t)}{\mathrm{d}t} &= T(n|n+1)P(n+1,t) \\ &\quad + T(n|n-1)P(n-1,t) \\ &\quad - \left[T(n-1|n) + T(n+1|n)\right]P(n,t)\,. \end{aligned} \tag{23}$$

For simplicity let us write

$$g_n = T(n+1|n) \quad \text{and} \quad r_n = T(n-1|n)\,, \tag{24}$$

then the master equation may be written as

$$\begin{aligned}\frac{\mathrm{d}P(n,t)}{\mathrm{d}t} &= r_{n+1}P(n+1,t) + g_{n-1}P(n-1,t) \\ &\quad - \left[r_n + g_n\right]P(n,t)\,. \end{aligned} \tag{25}$$

Examples of Master Equations

1. *The simple birth–death process*. For a population of simple organisms, for example a colony of bacteria, it might be reasonable to assume that the rate of birth of new bacteria is proportional to the number present at that time, and similarly for the rate of death. This is clearly a Markov process with $g_n = bn$ and $r_n = dn$, where b and d are rate constants. A variant is to include "immigrants" coming into the population from the outside at a constant rate c, so that $g_n = bn + c$.

   In this example g_n and r_n are linear in n. Such *linear one-step processes* can be solved by the introduction of the generating function $F(z, t) = \sum_n P(n, t)z^n$. This converts the master equation (which is a differential-difference equation) into a partial differential equation for $F(z, t)$ which can be solved if the process is linear. The simplest case of a pure death process ($b = c = 0$ in the above) can illustrate the general procedure. By rescaling the time ($t = \tau/d$), we may write the master equation in the very simple form

$$\frac{\mathrm{d}P(n,\tau)}{\mathrm{d}\tau} = (n+1)\,P(n+1,\tau) - n\,P(n,\tau)\,.$$

Multiplying this equation by z^n and summing over all $n \geq 0$ gives

$$\begin{aligned}&\frac{\partial}{\partial \tau}\left\{\sum_{n=0}^{\infty} z^n P(n,\tau)\right\}\\&= \sum_{n=0}^{\infty}(n+1)z^n P(n+1,\tau) - \sum_{n=1}^{\infty} n z^n P(n,\tau)\\&= \sum_{m=1}^{\infty} m z^{m-1} P(m,\tau) - z\sum_{n=1}^{\infty} n z^{n-1} P(n,\tau)\,,\end{aligned}$$

that is,

$$\frac{\partial F}{\partial \tau} = (1-z)\,\frac{\partial F}{\partial z}\,.$$

A change of variable to $\xi = (1-z)\mathrm{e}^{-\tau}$ and $\eta = \tau$ shows F to be a function of ξ only: $F(z,\tau) = \phi([1-z]\mathrm{e}^{-\tau})$. The function ϕ may be determined from the initial condition. For instance, if $P(n,0) = \delta_{n\,N}$, then $F(z,0) = z^N$ and so $\phi(\xi) = (1-\xi)^N$. This gives the solution for F to be

$$F(z,\tau) = \left[\left(1-\mathrm{e}^{-\tau}\right) + z\mathrm{e}^{-\tau}\right]^N\,.$$

In this case F can easily be expanded as a power series in z, and the $P(n,\tau)$ read off, but even if this is difficult, the moments of the distribution can be readily found by differentiation with respect to z and then setting $z = 1$. It should now be clear why F is called a generating function. In the general case the partial differential equation for F may be solved by standard methods [50]. The solution for a birth–death process without immigration is given in the book by Reichl [45]. The solution with immigration was first given by Kendall [25], who also introduced the technique of the generating function as a method of solution of the master equation.

2. *The Moran model of genetic drift*. Stochastic processes occur extensively in population genetics. The simplest, and most widely known, is a model of genetic drift introduced by Fisher [13] and Wright [58], in which a population of individuals in generation t mate randomly to produce the new generation $t+1$. We assume, for simplicity, that each individual has only one gene of a particular type, and that this may exist in one of two forms (alleles) denoted by A and B. The Wright–Fisher model is based on the sampling of the gene pool at generation t, which consists of n genes of type A and $(N-n)$ genes of type B, to produce the next generation of N genes. Although this may be formulated as a Markov chain, neither Fisher nor Wright did so; this was first carried out by Malécot [31] in 1944. Here we will describe a variant of the model introduced by Moran [36,37], which is a one-step process and can be formulated as a master equation.
The Moran model does not have non-overlapping generations, as in the Wright–Fisher model, and is more akin to a birth–death process where birth and death are coupled. At a given time, two individuals are sampled with replacement: one is designated the parent which is copied to create an offspring and the other is sacrificed to make way for the new offspring. Clearly if a B (chosen with probability $(N-n)/N$) is sacrificed and an A (chosen with probability n/N) is copied, this gives a contribution to $T(n+1|n)$. If the choice is that with A and B interchanged, this gives a contribution to $T(n-1|n)$. The transition rates for the Moran model are thus

$$\begin{aligned}T(n+1|n) &= \beta\left(1-\frac{n}{N}\right)\left(\frac{n}{N}\right)\,,\\T(n-1|n) &= \beta\left(\frac{n}{N}\right)\left(1-\frac{n}{N}\right)\,,\end{aligned}\qquad(26)$$

where β is a rate constant, which may be absorbed into the time t.
This may be extended in various ways. For example, mutations may be included: $A \overset{u}{\rightarrow} B$ and $B \overset{v}{\rightarrow} A$. With probability $(1-u-v)$ the offspring is taken to be a copy of the parent without mutation, as previously described. For the rest of the time (that is, with probability $u+v$), a mutation occurs. If the parent is an A, the offspring becomes a B with probability $u/(u+v)$, and if the parent is a B, the offspring becomes an A with probability $v/(u+v)$. This leads to the transition rates

$$\begin{aligned}T(n+1|n) &= (1-u-v)\left(1-\frac{n}{N}\right)\left(\frac{n}{N}\right)\\&\qquad + v\left(1-\frac{n}{N}\right)\,,\\T(n-1|n) &= (1-u-v)\left(1-\frac{n}{N}\right)\left(\frac{n}{N}\right)+u\left(\frac{n}{N}\right).\end{aligned}\qquad(27)$$

The master equation for the Moran model will be discussed again in the next section.

3. *Competition in a single species model*. The birth–death process described in Example 1 can be generalized to more complex ecological situations. As it stands it consists of the two processes $A \overset{d}{\rightarrow} E$ and $A \overset{b}{\rightarrow} A + A$ representing death and birth respectively. Here A represents an individual and E is a null state. To model the finite resources available in a given patch, we

put a limit on the number of allowed individuals: $n = 0, 1, \ldots, N$. We can also only allow a birth if enough space and/or other resources are available: $A + E \xrightarrow{b} A + A$ and include competition for these resources: $A + A \xrightarrow{c} A + E$. Since the probability of obtaining an A when sampling the patch is n/N and of obtaining an E is $(N - n)/N$, the birth term is now proportional to $n(N - n)/[N(N - 1)]$ and the competition term proportional to $n(n - 1)/[N(N - 1)]$. This gives the transition rates to be

$$\begin{aligned} T(n+1|n) &= \frac{2bn(N-n)}{N(N-1)}\,, \\ T(n-1|n) &= \frac{cn(n-1)}{N(N-1)} + \frac{dn}{N}\,. \end{aligned} \tag{28}$$

This approach can be extended to more than one species, for instance competition between and within two species [34] or predator-prey interactions [35]. In these cases the state space is multi-dimensional: $\mathbf{n} = (n_1, n_2, \ldots)$. The master equation still has the form (21), but with n replaced everywhere by the vector $\mathbf{n}$.

Whether or not a stationary state of the master equation exists depends on the nature of the boundary conditions. There are many types of boundary conditions, but two are particularly important. If the boundaries are reflecting, then the probability current vanishes there. If they are absorbing, then the probability of being at the boundary is zero. In the former case probability is conserved, in the latter case it is not, and leaks out of the system.

So to find a non-zero pdf as $t \to \infty$ (a stationary distribution) we therefore assume that the system lies within two reflecting boundaries. For a one-step process, the net flow of probability from the state n to the state $n + 1$, is $J(n, t) = g_n P(n, t) - r_{n+1} P(n + 1, t)$, where $J(n, t)$ is the probability current. The master Eq. (25) may be written as

$$\frac{\mathrm{d}P(n,t)}{\mathrm{d}t} = J(n-1,t) - J(n,t)\,.$$

For a stationary state, the left-hand side of this equation is zero, and the currents will be time-independent. Therefore $J(n - 1) = J(n)$ for all n, that is, all the currents are equal. Since the current vanishes at the boundaries, this constant must be zero. Therefore, for reflecting boundary conditions, $r_{n+1} P_{\mathrm{st}}(n + 1) = g_n P_{\mathrm{st}}(n)$ for all n. If we suppose that one boundary is at $n = 0$ and the other at $n = N$, then we have that $P_{\mathrm{st}}(1) = (g_0/r_1)P_{\mathrm{st}}(0)$, $P_{\mathrm{st}}(2) = (g_1/r_2)P_{\mathrm{st}}(1), \ldots,$ which implies $P_{\mathrm{st}}(2) = (g_1 g_0)/(r_2 r_1) P_{\mathrm{st}}(0), \ldots$ Iterating, the stationary state can be expressed as a simple product:

$$P_{\mathrm{st}}(n) = \frac{g_{n-1} g_{n-2} \ldots g_0}{r_n r_{n-1} \ldots r_1} P_{\mathrm{st}}(0)\,, \quad n = 1, \ldots, N\,. \tag{29}$$

The constant $P_{\mathrm{st}}(0)$ is determined by normalization:

$$\begin{aligned} \sum_{n=0}^{N} P_{\mathrm{st}}(n) &= P_{\mathrm{st}}(0) + \sum_{n>0} P_{\mathrm{st}}(n) = 1 \\ \Rightarrow (P_{\mathrm{st}}(0))^{-1} &= 1 + \sum_{n=1}^{N} \frac{g_{n-1} g_{n-2} \ldots g_0}{r_n r_{n-1} \ldots r_1}\,. \end{aligned} \tag{30}$$

As an example, we return to the Ehrenfest urn (12), which in the language of the master equation is defined by $g_n = (N - n)/N$ and $r_n = n/N$ (any overall rate may be absorbed into the time, and this is irrelevant as far as the stationary state is concerned). Here $n = 0, 1, \ldots, N$ and the molecules never go outside this range, so the boundaries are reflecting. Applying Eqs. (29) and (30) shows that the stationary state is the binomial distribution given by Eq. (15).

The Fokker–Planck Equation

The Fokker–Planck equation describes stochastic processes at a more coarse grained level than those that we have discussed so far. It only involves continuous stochastic variables; these could be for instance the fraction of individuals or genes of a certain kind in a population, whereas the master equation recognized the individuals or genes as discrete entities. To obtain the Fokker–Planck equation we first derive the Kramers–Moyal expansion [29,38].

We begin by defining the *jump moments* for the system:

$$M_\ell(x, t, \Delta t) = \int \mathrm{d}\xi \;(\xi - x)^\ell \, P(\xi, t + \Delta t | x, t)\,. \tag{31}$$

We will assume that these are known, that is, they can be obtained by some other means. They will, however, only be required in the limit of small Δt.

The starting point for the derivation is the Chapman–Kolmogorov equation (4), with the choice of variables analogous to that used in Eq. (17) for the discrete case:

$$P(x, t + \Delta t) = \int \mathrm{d}x' P(x, t + \Delta t | x', t)\, P(x', t)\,, \tag{32}$$

again dropping the dependence on the initial conditions. The integrand may be written as

$$\begin{aligned} & P(x,t+\Delta t|x',t)\,P(x',t) \\ & = P([x-\Delta x]+\Delta x, t+\Delta t|[x-\Delta x],t)P([x-\Delta x],t) \\ & = \sum_{\ell=0}^{\infty}\frac{(-1)^{\ell}}{\ell!}(\Delta x)^{\ell} \\ & \qquad \cdot \frac{\partial^{\ell}}{\partial x^{\ell}}\{P(x+\Delta x, t+\Delta t|x,t)P(x,t)\}\ , \end{aligned} \tag{33}$$

where $\Delta x = x - x'$. Integrating over x' gives for Eq. (32):

$$P(x,t+\Delta t) = \sum_{\ell=0}^{\infty}\frac{(-1)^{\ell}}{\ell!}\frac{\partial^{\ell}}{\partial x^{\ell}}\{M_{\ell}(x,t,\Delta t)P(x,t)\}\ . \tag{34}$$

Since $P(\xi,t|x,t) = \delta(\xi - x)$, it follows from Eq. (31), that $\lim_{\Delta t\to 0} M_{\ell}(x,t,\Delta t) = 0$ for $\ell \geq 1$. Also $M_0(x,t,\Delta t) = 1$. Bearing these results in mind, we will now assume that the jump moments for $\ell \geq 1$ take the form

$$M_{\ell}(x,t,\Delta t) = D^{(\ell)}(x,t)\Delta t + o(\Delta t)\ . \tag{35}$$

Substituting this into Eq. (34), dividing by Δt and taking the limit $\Delta t \to 0$ gives

$$\frac{\partial P}{\partial t} = \sum_{\ell=1}^{\infty}\frac{(-1)^{\ell}}{\ell!}\frac{\partial^{\ell}}{\partial x^{\ell}}\left\{D^{(\ell)}(x,t)P(x,t)\right\}\ . \tag{36}$$

Equation (36) is the Kramers–Moyal expansion. So far nothing has been assumed other than the Markov property and the existence of Taylor series expansions. However, in many situations, examination of the jump moments reveal that in a suitable approximation they may be neglected for $\ell > 2$. In this case, we may truncate the Kramers–Moyal expansion (36) at second order and obtain the *Fokker–Planck equation*:

$$\frac{\partial P}{\partial t} = -\frac{\partial}{\partial x}\left[A(x,t)P(x,t)\right] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[B(x,t)P(x,t)\right]\ , \tag{37}$$

where $A = D^{(1)}$ and $B = D^{(2)}$ are independent of t if the process is stationary.

To calculate the jump moments (31), it is convenient to write them in terms of the underlying stochastic process $x(t)$. We use the notation

$$\langle x(t)\rangle_{x(t_0)=x_0} = \int \mathrm{d}x\, xP(x,t|x_0,t_0)\,, \tag{38}$$

for the mean of the stochastic variable at time t, conditional on the value of $x(t)$ being given to be x_0 at time t_0. With this notation $x(t)$ denotes the process and the angle brackets are averages over realizations of this process. More generally, we may define $\langle f(x(t))\rangle$ in a similar way, and in particular the jump moments are given by

$$M_{\ell}(x,t,\Delta t) = \langle (x(t+\Delta t)-x)^{\ell}\rangle_{x(t)=x}\ . \tag{39}$$

Examples of Fokker–Planck Equations

1. *Simple diffusion*. For the simple symmetric random walk, $g_n = 1$ and $r_n = 1$ when expressed in the language of the master equation (after a rescaling of the time so that the rates may taken to be equal to unity). From Eqs. (18) and (19) we find that for a one-step stationary process,

$$\begin{aligned} & \langle (n(t+\Delta t)-n)^{\ell}\rangle_{n(t)=n} \\ & = \begin{cases} (g_n - r_n)\,\Delta t + o\,(\Delta t)\ , & \text{if } \ell \text{ is odd} \\ (g_n + r_n)\,\Delta t + o\,(\Delta t)\ , & \text{if } \ell \text{ is even}\,, \end{cases} \end{aligned} \tag{40}$$

   and so for the symmetric random walk the odd moments all vanish, and the even moments are equal to $2\Delta t + o(\Delta t)$. We now make the approximation which will yield the Fokker–Planck equation: we let $x = nL$, where L is the step size, and let $L \to 0$. Since, for ℓ even, $\langle (x(t+\Delta t)-x)^{\ell}\rangle = (2L^{\ell})\Delta t$, if we rescale the time by introducing $\tau = L^2 t$, then all jump moments higher than the second disappear in the limit $L \to 0$, and Eq. (36) becomes

$$\frac{\partial P}{\partial t} = \frac{\partial^2 P}{\partial x^2}\ . \tag{41}$$

   This is the familiar diffusion equation obtained from a continuum approximation to the discrete random walk.
2. *The diffusion limit of the Moran model*. For the Moran model with no mutation, we have from Eqs. (26) and (40) that the odd moments again vanish. If we describe the process by $x(t) = n(t)/N$, the fraction of the genes that are of type A at time t, then the even jump moments are given by

$$\begin{aligned} & \left\langle (x(t+\Delta t)-x)^{\ell}\right\rangle_{x(t)=x} \\ & \qquad = \frac{1}{N^{\ell}}2x\,(1-x)\,\Delta t + o\,(\Delta t)\ , \end{aligned}$$

   and so introducing a rescaled time $\tau = 2t/N^2$, and letting $N \to \infty$ we obtain the Fokker–Planck equation

$$\frac{\partial P}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[x(1-x)P\right]\ . \tag{42}$$

The factor of 2 in the rescaling of the time is included so that the diffusive form of the Moran model agrees with that found in the Wright–Fisher model [6].

Now suppose mutations are included. The transition rates are given by Eq. (27) and lead to jump moments $\langle (x(t+\Delta t)-x)^{\ell}\rangle \sim N^{-\ell}$. So the first and second jump moments are not of the same order, and the introduction of a rescaled time $\tau = t/N$, and subsequently letting $N \to \infty$, gives a Fokker–Planck equation of the form (37), but with $B = 0$ and $A(x) = v - (u+v)x$. This corresponds to a *deterministic* process with $\mathrm{d}x/\mathrm{d}t = v - (u+v)x$ [22]. The fact that the system tends to a macroscopic equation when $N \to \infty$ has to be taken into account when determining the nature of the fluctuations for large N. We will discuss this further in Sect. "System Size Expansion".

On the other hand, suppose that the mutation rates scale with N according to $u = 2\tilde{u}/N$ and $v = 2\tilde{v}/N$, where $\tilde{u}$ and $\tilde{v}$ have a finite limit as $N \to \infty$, and where the 2 has again been chosen to agree with the Wright–Fisher model. Now both the first and second jump moments are of order N^{-2}, with the higher moments falling off faster with N. Therefore once again introducing the rescaled time $\tau = 2t/N^2$ and letting $N \to \infty$, we obtain the Fokker–Planck equation

$$\frac{\partial P}{\partial t} = -\frac{\partial}{\partial x}\left[\{\tilde{v} - (\tilde{u}+\tilde{v})x\}P\right] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[x(1-x)P\right] . \tag{43}$$

So depending on the precise scaling with N, the Moran model gives different limits as $N \to \infty$ [24]. In the first case, the mutational effects are strong enough that a macroscopic description exists, with fluctuations about the macroscopic state, as discussed in Sect. "System Size Expansion". In the second case, the mutational effects are weaker, and there is no macroscopic equation describing the system; the large N limit is a nonlinear Fokker–Planck equation of the diffusive type.

The Fokker–Planck equation (37) for stationary processes, where A and B are functions of x only, can be solved by separation of variables, with solutions of the form $p(x)\mathrm{e}^{-\lambda t}$, where λ is a constant. The equation for $p(x)$ is then a second order differential equation which, when boundary conditions are given, is of the Sturm–Liouville type. To specify the boundary conditions we once again introduce the probability current, this time through the continuity equation,

$$\frac{\partial P(x,t)}{\partial t} + \frac{\partial J(x,t)}{\partial x} = 0 , \tag{44}$$

where the probability current $J(x,t)$ is given by

$$J(x,t) = A(x,t)P(x,t) - \frac{1}{2}\frac{\partial}{\partial x}\left[B(x,t)P(x,t)\right] . \tag{45}$$

Let us suppose that the system is defined on the interval $[a,b]$. Then if the boundaries are reflecting, there is no net flow of probability across $x = a$ and $x = b$. This implies that $J(a,t) = 0$ and $J(b,t) = 0$. If we integrate the equation of continuity (44) from $x = a$ to $x = b$, and apply these boundary conditions, we see that $\int_a^b P(x,t)\mathrm{d}x$ is independent of time. Therefore, if the pdf is initially normalized, it remains normalized. This is in contrast with the case of absorbing boundary conditions, defined by $P(a,t) = 0$ and $P(b,t) = 0$. If the boundary conditions are at infinity we require that $\lim_{x\to\pm\infty} P(x,t) = 0$, so that if P is well-behaved it is normalizable, and also that $\partial P/\partial x$ is well-behaved in this limit: $\lim_{x\to\pm\infty} \partial P/\partial x = 0$. If A or B do not diverge as $x \to \pm\infty$ this implies that $\lim_{x\to\pm\infty} J(x,t) = 0$. Other types of boundary conditions are possible, and we do not attempt a complete classification here [16,46].

If A and B are independent of time, then from Eq. (44), the stationary state of the system must be given by $\mathrm{d}J(x)/\mathrm{d}x = 0$, that is, J is a constant. For reflecting boundary conditions this constant is zero, and so from Eq. (45) the stationary pdf, $P_{\mathrm{st}}(x)$ must satisfy

$$0 = A(x)P_{\mathrm{st}}(x) - \frac{1}{2}\frac{\partial}{\partial x}\left[B(x)P_{\mathrm{st}}(x)\right] . \tag{46}$$

This may be integrated to give

$$P_{\mathrm{st}}(x) = \frac{C}{B(x)}\exp\left\{2\int^x \mathrm{d}x' \quad \frac{A(x')}{B(x')}\right\} , \tag{47}$$

where C is a constant which has to be chosen so that $P_{\mathrm{st}}(x)$ is normalized.

The Fokker–Planck equation, with A independent of t and B constant, can be transformed into the Schrödinger-like problem

$$-B\frac{\partial \psi}{\partial t} = -\frac{B^2}{2}\frac{\partial^2 \psi}{\partial x^2} + U(x)\psi , \tag{48}$$

by the transformation

$$P(x,t) = \left[P_{\mathrm{st}}(x)\right]^{1/2}\psi(x,t) , \tag{49}$$

where

$$U(x) = \frac{1}{2}\left[A(x)\right]^2 + \frac{B}{2}\frac{\mathrm{d}A}{\mathrm{d}x} . \tag{50}$$

So a one-dimensional stationary stochastic process, under certain conditions (such as constant second jump moment) is equivalent to quantum mechanics in imaginary time, with B taking over the role of Planck's constant.

As an example we consider the *Ornstein–Uhlenbeck process* defined by

$$\frac{\partial P}{\partial t} = \frac{\partial}{\partial x}[axP] + D\frac{\partial^2 P}{\partial x^2}, \quad x \in (-\infty, \infty), \quad a > 0. \quad (51)$$

In this case, the potential (50) is given by $U(x) = [a^2x^2 - 2aD]/2$, and so the problem is equivalent to the one-dimensional simple harmonic oscillator in quantum mechanics, but with an energy shift. As in that problem [47], the eigenfunctions are Hermite polynomials. Specifically, the right-eigenfunctions are

$$p_m(x) = P_{\text{st}}(x)\frac{1}{[2^m m!]^{1/2}} H_m(\alpha x), \quad (52)$$

where H_m are the Hermite polynomials [1] and $\alpha = (a/2D)^{1/2}$. The eigenvalue corresponding to the eigenfunction (52) is $\lambda_m = am$, $m = 0, 1, \ldots$ and the left-eigenfunctions are $q_m(x) = [P_{\text{st}}(x)]^{-1} p_m(x)$. From these explicit solutions we may calculate other quantities of interest, such as correlation functions. We note that the eigenvalues are all non-negative and that the stationary state corresponds to $\lambda = 0$. The left-eigenfunction for the stationary state is equal to 1. These latter results hold true for a wide-class of such problems.

Stochastic Differential Equations

So far we have described stochastic processes in terms of equations which give the time evolution of pdfs. In this section, we will describe equations for the stochastic variables themselves. The most well known instance of such an equation is the Langevin equation for the velocity of a Brownian particle, and so we begin with this particular example.

Suppose that a small macroscopic particle of mass m (such as a pollen grain) is immersed in a liquid at a temperature T. In addition to any macroscopic motion that the particle may have, its velocity fluctuates due to the random collisions of the particle with the molecules of the liquid. For simplicity, we confine ourselves to one-dimensional motion – along the x-axis. Then the equation of motion of the particle may be written in the form

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = -\alpha\frac{\mathrm{d}x}{\mathrm{d}t} - \frac{\mathrm{d}V}{\mathrm{d}x} + \mathcal{F}(t). \quad (53)$$

The first term on the right-hand side is due to the viscosity of the fluid and α is the friction constant. The second term, where $V(x)$ is a potential, represents the interaction of the particle with any external forces, such as gravity. The final term is the random force due to collisions with the molecules of the liquid. Clearly to complete the specification of the dynamics of the particle we need to give (i) the initial position and velocity of the particle, and (ii) the statistics of the random force $\mathcal{F}(t)$.

To make progress with these points, we imagine a large number of realizations of the dynamics, in which the particle starts with the same initial position, x_0, and velocity, v_0, but where the initial positions and velocities of the molecules in the liquid will be different. Taking the average over a large number of such realizations will give the average position $\langle x(t)\rangle$ and velocity $\langle v(t)\rangle$ at time t, conditional on $x(0) = x_0$ and $v(0) = v_0$. The statistics of the fluctuating force $\mathcal{F}(t)$ are assumed to be such that

(a) $\langle \mathcal{F}(t)\rangle = 0$, since we do not expect one direction to be favored over the other.

(b) $\langle \mathcal{F}(t)\mathcal{F}(t')\rangle = 2D\delta(t - t')$, since we expect that after a few molecular collisions the value that $\mathcal{F}$ takes on will be independent of its former value. That is, the force $\mathcal{F}$ becomes uncorrelated over times of the order of a few collision times between molecules. This is tiny on observational time scales, and so taking the correlation function to be a delta-function is an excellent approximation. The weight of the delta-function is denoted by $2D$, where at this stage D is undetermined.

(c) $\mathcal{F}(t)$ is taken to be Gaussianly distributed on grounds of simplicity, but also because by the central limit theorem it is assumed that the net effect of the large number of molecules which collide with the pollen grain will lead to a distribution which is Gaussian.

Since a Gaussian distribution is specified by its first two moments, conditions (a), (b) and (c) completely define the statistics of $\mathcal{F}(t)$.

Finally, Eq. (53) as it stands does not define a Markov process. This is most easily seen if we write down a discrete time version of the equation. The second derivative means that $x(t + \delta t)$ not only depends on $x(t)$, but also on $x(t - \delta t)$. Therefore only first order derivatives should be included in such equations if the process is to be Markov. This is easily achieved by promoting $v(t)$ to be a second stochastic variable in addition to $x(t)$. Then Eq. (53) may be equivalent written as

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= v\,, \\ m\frac{\mathrm{d}v}{\mathrm{d}t} &= -\alpha v - \frac{\mathrm{d}V}{\mathrm{d}x} + \mathcal{F}(t)\,, \end{aligned} \quad (54)$$

which does define a Markov process. Although, as we remarked in the Introduction, we deal almost exclusively with Markov processes in this article, the situation we have just discussed is a good illustration of one way of dealing with processes which are presented as being non-Markovian. The method simply consists of adding a sufficient number of supplementary variables to the definition of the state variables of the process until it becomes Markovian. There is no guarantee that this will be possible or require only a small number of additional variables to be promoted in this way, but it is the most straightforward and direct way of rendering non-Markovian processes tractable.

To begin the analysis of Eq. (54) we assume that there are no external forces and so the term $\mathrm{d}V/\mathrm{d}x$ is equal to zero. We may then write Eq. (54) as the Langevin equation

$$\frac{\mathrm{d}v}{\mathrm{d}t} = -\gamma v + F(t)\,; \quad v(0) = v_0\,, \tag{55}$$

where $\gamma = \alpha/m$ and $F(t) = \mathcal{F}(t)/m$. This implies that

$$\langle F(t)\rangle = 0 \ \text{ and } \ \langle F(t)F(t')\rangle = \frac{2D}{m^2}\delta(t-t')\,. \tag{56}$$

Note that since $F(t)$ is a random variable, solving the Langevin equation will give $v(t)$ as a random variable (having a known distribution). It is therefore a *stochastic differential equation*. The function F is frequently called "the noise term" or simply "the noise". It is *white noise* since the Fourier transform of a delta-function is a constant – all frequencies are present in equal amounts.

Multiplying the Langevin equation (55) by the integrating factor $\mathrm{e}^{\gamma t}$ gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[v(t)\mathrm{e}^{\gamma t}\right] = F(t)\mathrm{e}^{\gamma t} \Rightarrow v(t) = v_0\mathrm{e}^{-\gamma t} + \mathrm{e}^{-\gamma t}\int_0^t \mathrm{d}t'\,F(t')\mathrm{e}^{\gamma t'}\,. \tag{57}$$

By taking the average of the expression for $v(t)$ we find $\langle v(t)\rangle = v_0\mathrm{e}^{-\gamma t}$. More interestingly, if we square the expression for $v(t)$ and take the average, then we find

$$\langle v^2(t)\rangle = v_0^2\mathrm{e}^{-2\gamma t} + \frac{D}{\alpha m}\left[1-\mathrm{e}^{-2\gamma t}\right]\,, \tag{58}$$

which implies that

$$\lim_{t\to\infty}\langle v^2(t)\rangle = \frac{D}{\alpha m}\,. \tag{59}$$

On the other hand, as $t\to\infty$, the Brownian particle will be in thermal equilibrium:

$$\lim_{t\to\infty}\langle v^2(t)\rangle = v_{\mathrm{eq}}^2 \quad \text{and} \quad \frac{1}{2}mv_{\mathrm{eq}}^2 = \frac{1}{2}kT\,,$$

where T is the temperature of the liquid and k is Boltzmann's constant. This implies that

$$\frac{1}{2}m\left(\frac{D}{\alpha m}\right) = \frac{1}{2}kT \ \Rightarrow \ D = \alpha kT\,. \tag{60}$$

The molecules of the liquid are acting as a heat bath for the system – which in this case is a single Brownian particle. The equation $D = \alpha kT$ is a simple example of a fluctuation-dissipation theorem, and determines D in terms of the friction constant, α, and of the temperature of the liquid, T.

Although we have presented a somewhat heuristic rationale for Eq. (54), it may be derived in a more controlled way. A particularly clear derivation has been given by Zwanzig [60], where the starting point is a Hamiltonian which contains three terms: for the system, the heat bath and the interaction between the system and the heat bath. Taking the heat bath to be made up of coupled harmonic oscillators and the interaction term between the system and heat bath to be linear, it is possible to integrate out the bath degrees of freedom exactly, and be left only with the equations of motion of the system degrees of freedom plus the initial conditions of the bath degrees of freedom. Assuming that the bath is initially in thermal equilibrium, so that these initial values are distributed according to a Boltzmann distribution, adds extra "noise" terms to the equations of motion which, with a few more plausible assumptions, make them of the Langevin type.

Examples of Langevin–like Equations

1. *Overdamped Brownian motion*. Frequently the viscous damping force $-\alpha v$ is much larger than the inertial term $m\mathrm{d}^2x/\mathrm{d}t^2$ in Eq. (53), and so to a good approximation the left-hand side of Eq. (53) can be neglected. Scaling time by α, we arrive at the Langevin equation for the motion of an overdamped Brownian particle:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -V'(x) + \mathcal{F}(t)\,; \quad x(0) = x_0\,, \tag{61}$$

where $'$ denotes differentiation with respect to x and where, due to the rescaling of time by α,

$$\langle\mathcal{F}(t)\rangle = 0; \quad \langle\mathcal{F}(t)\mathcal{F}(t')\rangle = 2\tilde{D}\delta(t-t'); \quad \tilde{D} = \frac{D}{\alpha}\,. \tag{62}$$

A particularly well-known case is when the Brownian particle is moving in the harmonic potential $V(x) = ax^2/2$. Then

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -ax + \mathcal{F}(t)\,; \quad x(0) = x_0\,. \tag{63}$$

Since Eq. (63) relating $x(t)$ to $\mathcal{F}(t)$ is linear, and since the distribution of $\mathcal{F}(t)$ is Gaussian, then $x(t)$ is also distributed according to a Gaussian distribution. Comparing with Eqs. (55) and (56), which also define a linear system, we find that $\langle x(t)\rangle = x_0\mathrm{e}^{-at}$ and, from

Eq. (58), $\langle x^2(t)\rangle = x_0^2 \mathrm{e}^{-2at} + (\tilde{D}/a)\left[1-\mathrm{e}^{-2at}\right]$. This gives

$$P(x,t|x_0,0) = \sqrt{\frac{a}{2\pi\tilde{D}\left[1-\mathrm{e}^{-2at}\right]}}\exp\left\{-\frac{a\left(x-x_0\mathrm{e}^{-at}\right)^2}{2\tilde{D}\left[1-\mathrm{e}^{-2at}\right]}\right\} . \quad (64)$$

It is straightforward to check that this conditional pdf satisfies the Fokker–Planck equation (51) for the Ornstein–Uhlenbeck process. Below we will show this more directly, by starting from the Langevin equation (61) and deriving Eq. (51) if $V(x)$ is quadratic.

Another case of interest is when $V(x)$ is a double-well potential, $V(x) = -ax^2/2 + bx^4/4$. If the particle is initially located near the bottom of one of the potential wells, it will take on average a time of the order of $\mathrm{e}^{\Delta V/D}$ to hop over the barrier and into the well on the other side. Here ΔV is the height of the barrier that it has to hop over [29].

2. *Environmental noise in population biology*. One of the simplest models of two species with population sizes N_1 and N_2 which are competing for a common resource, is the two coupled deterministic ordinary differential equations $\dot{N}_i = r_i N_i$, $i = 1,2$. The growth rates, r_i, depend on the population sizes in such a way that as the population sizes increase, the r_i decrease to reflect the increased competition for resources. This could be modeled, for instance, by taking $r_i = a_i - b_{ii}N_i - b_{ij}N_j$ with $i,j = 1,2$ and $i \neq j$. In reality, external factors such as climate, terrain, the presence of other species, and indeed any factor which has an uncertain influence on these two species, will also affect the growth rate. This can be modeled by adding an external random term to the r_i which represents this *environmental stochasticity* [33]. Then the equations become

$$\begin{aligned}\frac{\mathrm{d}N_1}{\mathrm{d}t} &= a_1N_1 - b_{11}N_1^2 - b_{12}N_1N_2 + N_1\zeta_1(t)\\ \frac{\mathrm{d}N_2}{\mathrm{d}t} &= a_2N_2 - b_{22}N_2^2 - b_{21}N_2N_1 + N_2\zeta_2(t) .\end{aligned} \quad (65)$$

Since the noise terms, $\zeta_i(t)$ are designed to reflect the large number of coupled variables omitted from the description of the model, it is natural, by virtue of the central limit theorem, to assume that they are Gaussianly distributed. It also seems reasonable to assume that any temporal correlation between these external influences is on scales very much shorter than those of interest to us here, and that the noises have zero mean. We therefore assume that

$$\langle\zeta_i(t)\rangle = 0 ; \quad \langle\zeta_i(t)\zeta_j(t')\rangle = 2D_i\delta_{ij}\delta(t-t') , \quad (66)$$

where the D_i describe the strength of the stochastic effects. The deterministic equations (that is, Eq. (65) without the noise terms) have a fixed point at the origin, one on each of the N_1 and N_2 axes, and may have another at non-zero N_1 and N_2. For some values of the parameters this latter fixed point may be a saddle, with those on the axes being stable and the origin unstable. In this situation the eventual fate of the species depends significantly on the noise: if the combination of the nonlinear dynamics and the noise drives the system to the vicinity of the fixed point on the N_1 axis, then species 2 will become extinct, and vice-versa.

Langevin equations with Gaussian white noise are equivalent to Fokker–Planck equations. This can be most easily seen by calculating the jump moments (39) from the Langevin equation. For instance, if we begin from the Langevin equation for an overdamped Brownian particle (61),

$$\begin{aligned}\Delta x(t) \equiv x(t+\Delta t) - x(t) &= \int_t^{t+\Delta t} \mathrm{d}t'\, \dot{x}(t')\\ &= -\int_t^{t+\Delta t} \mathrm{d}t'\, V'(x(t')) + \eta(t) ,\end{aligned} \quad (67)$$

where $\eta(t) = \int_t^{t+\Delta t} \mathrm{d}t'\, \mathcal{F}(t')$. From Eq. (62) it is straightforward to calculate the moments of $\eta(t)$: $\langle\eta(t)\rangle = 0$,

$$\langle\eta^2(t)\rangle = \int_t^{t+\Delta t} \mathrm{d}t' \int_t^{t+\Delta t} \mathrm{d}t'' \langle\mathcal{F}(t')\mathcal{F}(t'')\rangle = 2\tilde{D}\Delta t \quad (68)$$

and, since $\eta(t)$ is Gaussian, $\langle\eta^n(t)\rangle$ is zero if n is odd, and at least of order $(\Delta t)^2$ for $n \geq 4$. This implies that

$$\begin{aligned}M_1(x,\Delta t) &= -V'(x)\Delta t + \mathcal{O}\left(\Delta t\right)^2 ,\\ M_2(x,\Delta t) &= 2\tilde{D}\Delta t + \mathcal{O}\left(\Delta t\right)^2 ,\end{aligned} \quad (69)$$

with all moments of order $(\Delta t)^2$ or higher for $\ell > 2$. The notation $\mathcal{O}\left(\Delta t\right)^2$ means that the magnitude of this quantity is less than a constant times $(\Delta t)^2$, for sufficiently small nonzero $(\Delta t)^2$. This is a weaker, but more specific, statement than saying it is $o\left(\Delta t\right)$. Using Eqs. (35) and (36), the Fokker–Planck equation which is equivalent to the

Langevin equation (61) is found to be

$$\frac{\partial P}{\partial t} = \frac{\partial}{\partial x}\left[V'(x)P\right] + \tilde{D}\frac{\partial^2 P}{\partial x^2} . \tag{70}$$

From Eq. (47), the stationary pdf is $P_{\text{st}}(x) = C\exp\{-V(x)/\tilde{D}\} = C\exp\{-V(x)/kT\}$, as expected.

Although in this article we have largely restricted our attention to stochastic processes involving one variable, the construction of a Fokker–Planck equation from the Langevin equation goes through in a similar way for an n-dimensional process $\mathbf{x} = (x_1, \dots, x_n)$. In this case the jump moments are

$$\left\langle \Delta x_{i_1}(t)\Delta x_{i_2}(t)\dots\Delta x_{i_\ell}(t)\right\rangle_{\mathbf{x}(t)=\mathbf{x}} = D_{i_1\dots i_\ell}(\mathbf{x},t)\Delta t + o(\Delta t) , \tag{71}$$

where $\Delta x_{i_\alpha} = x_{i_\alpha}(t+\Delta t) - x_{i_\alpha}$. The Kramers–Moyal expansion is then

$$\frac{\partial P}{\partial t} = \sum_{\ell=1}^{\infty}\frac{(-1)^\ell}{\ell!}\frac{\partial^\ell}{\partial x_{i_1}\dots\partial x_{i_\ell}}\left\{D_{i_1\dots i_\ell}(\mathbf{x},t)P\right\} . \tag{72}$$

The Langevin equation for Brownian motion (54), without going to the overdamped limit, serves as a simple illustration of this generalization. Here $\Delta x(t) = v\Delta t$ and $\Delta v(t) = -\gamma v\,\Delta t - m^{-1}V'(x)\,\Delta t + m^{-1}\eta(t)$. This results in the Fokker–Planck equation

$$\frac{\partial P}{\partial t} = -\frac{\partial}{\partial x}\left[vP\right] + \frac{\partial}{\partial v}\left[\left\{\gamma v + m^{-1}V'(x)\right\}P\right] + \frac{\gamma kT}{m}\frac{\partial^2 P}{\partial v^2} . \tag{73}$$

This is Kramer's equation. It has a stationary pdf $P_{\text{st}}(x,v) = C\exp\{-E/kT\}$, where $E = mv^2/2 + V(x)$.

We end this section by finding the Fokker–Planck equation which is equivalent to the general set of Langevin equations of the form

$$\dot{x}_i = A_i(\mathbf{x},t) + \sum_{\alpha=1}^{m} g_{i\alpha}(\mathbf{x},t)\,\zeta_\alpha(t); \quad i = 1,\dots,n, \tag{74}$$

where $\zeta_\alpha(t)$, $\alpha = 1,\dots,m$, is a Gaussian white noise with zero mean and with

$$\langle\zeta_\alpha(t)\zeta_\beta(t')\rangle = \delta_{\alpha\beta}\delta(t-t') . \tag{75}$$

Proceeding as in Eq. (67), but noting the dependence of the function $g_{i\alpha}$ on the stochastic variable, yields

$$\begin{aligned} M_i(\mathbf{x},t,\Delta t) &= \left[A_i(\mathbf{x},t) + \theta(0)\sum_{j=1}^{n}\sum_{\alpha=1}^{m} g_{j\alpha}(\mathbf{x},t)\frac{\partial}{\partial x_j}g_{i\alpha}(\mathbf{x},t)\right]\Delta t + \mathcal{O}(\Delta t)^2 , \\ M_{ij}(\mathbf{x},t,\Delta t) &= \sum_{\alpha=1}^{m}\left[g_{i\alpha}(\mathbf{x},t)g_{j\alpha}(\mathbf{x},t)\right]\Delta t + \mathcal{O}(\Delta t)^2 , \end{aligned} \tag{76}$$

with all jump moments higher than the second being of order $(\Delta t)^2$ or higher. The quantity $\theta(0)$ is the value of the Heaviside theta function, $\theta(x)$, at $x = 0$ and is indeterminate. This indicates that the Langevin description does not correspond to a unique Fokker–Planck equation. This situation occurs whenever the white noise in a Langevin equation is multiplied by a function which depends on the state variable, as in Eq. (74). For systems such as this acted upon by *multiplicative noise* the Langevin description has to be supplemented by a rule which says whether the state variable in the multiplying function ($g_{i\alpha}$ in Eq. (74)) is that before or after the noise pulse acts [52]. If it is taken to be the value immediately after the noise pulse acts then $\theta(0) = 0$ (Itô rule), whereas if it is taken to be the average of the values before and after, then $\theta(0) = 1/2$ (Stratonovich rule). The Fokker–Planck equation is now found from Eq. (72) to be

$$\frac{\partial P}{\partial t} = -\sum_{i=1}^{n}\frac{\partial}{\partial x_i}\left[A_i(\mathbf{x},t)P(\mathbf{x},t)\right] + \frac{1}{2}\sum_{i,j=1}^{n}\sum_{\alpha=1}^{m}\frac{\partial^2}{\partial x_i\partial x_j}\left[g_{i\alpha}(\mathbf{x},t)g_{j\alpha}(\mathbf{x},t)P(\mathbf{x},t)\right] , \tag{77}$$

in the Itô case and

$$\frac{\partial P}{\partial t} = -\sum_{i=1}^{n}\frac{\partial}{\partial x_i}\left[A_i(\mathbf{x},t)P(\mathbf{x},t)\right] + \frac{1}{2}\sum_{i,j=1}^{n}\sum_{\alpha=1}^{m}\frac{\partial}{\partial x_i}\left[g_{i\alpha}(\mathbf{x},t)\frac{\partial}{\partial x_j}\left\{g_{j\alpha}(\mathbf{x},t)P(\mathbf{x},t)\right\}\right], \tag{78}$$

in the Stratonovich case.

Path Integrals

While most early work on stochastic processes was concerned with linear systems, naturally attention soon moved on to the many interesting systems which could be modeled as nonlinear stochastic processes. These systems are much more difficult to analyze. For example, a nonlinear Langevin equation cannot be solved directly, and so the averaging procedure cannot be carried out in the same explicit way as described in Sect. "Stochastic Differential Equations". There is however one method which is applicable to many nonlinear stochastic differential equations of interest: the solution of these equations can be formally written down as a path-integral, and from this correlation functions and other quantities of physical interest can be obtained. This also has the advantage that all the formalism and approximation schemes developed to study functional integrals over the years can be called into play.

Path-integrals are intimately related to Brownian motion and the earliest work on the subject by Wiener [56, 57], emphasized this. If the problem of interest is formulated as a set of Langevin equations, the derivation of the path-integral representation is particularly straightforward, if rather heuristic. For clarity we begin with the simplest case: an overdamped system with a single degree of freedom, x, acted upon by white noise. The Langevin equation is given by Eq. (61) and the noise is defined by Eq. (62). Since the noise is assumed to be Gaussian, Eq. (62) is a complete specification. An equivalent way of giving it is through the pdf [12]:

$$P[\mathcal{F}]\,\mathcal{D}\mathcal{F} \propto \exp\left(-\frac{1}{4\tilde{D}}\int \mathrm{d}t \mathcal{F}^2(t)\right)\,\mathcal{D}\mathcal{F}\,, \tag{79}$$

where $\mathcal{D}\mathcal{F}$ is the functional measure. The idea is now to regard the Langevin equation (61) as defining a mapping $\mathcal{F} \mapsto x$. The pdf for the x variable is then given by

$$\begin{aligned} P[x] &= P[\mathcal{F}]|_{\mathcal{F}=\dot{x}+V'(x)}\,J[x] \\ &\propto \exp\left(-\frac{1}{4\tilde{D}}\int \mathrm{d}t\,[\dot{x}+V'(x)]^2\right)\,J[x]\,, \end{aligned} \tag{80}$$

where

$$J[x] = \det\left[\frac{\delta\mathcal{F}}{\delta x}\right]\,, \tag{81}$$

is the Jacobian of the transformation. An explicit expression for the Jacobian may be obtained either by direct calculation of a discretized form of the Langevin equation [21] or through use of the identity relating the determinant of a matrix to the exponential of the trace of the logarithm of that matrix [59]. One finds that $J[x] \propto \exp\{\theta(0)\int \mathrm{d}t\,V''(x)\}$. The quantity $\theta(0)$ is once again the indeterminate value of the Heaviside theta function $\theta(x)$ at $x = 0$. Its appearance is a reflection of the fact that, due to the Brownian-like nature of the paths in the functional integral, the nature of the discretization appears explicitly through this factor [48]. If we consistently use the mid-point rule throughout, then we may take $\theta(0) = 1/2$, which gives

$$\begin{aligned} P[x] &\propto \exp\left(-\frac{1}{4\tilde{D}}\int \mathrm{d}t\,[\dot{x}+V'(x)]^2 + \frac{1}{2}\int \mathrm{d}t\,V''(x)\right) \\ &= \exp\left(-S[x]/\tilde{D}\right)\,. \end{aligned} \tag{82}$$

All quantities of interest can now be found from expression (82). For example, the conditional probability distribution, $P(x,t|x_0,t_0)$ is given by

$$\langle\delta(x-x(t))\rangle_{x(t_0)=x_0} = \int_{x(t_0)=x_0} \mathcal{D}x\delta(x-x(t))\,P[x]\,. \tag{83}$$

The expression (82) has much in common with Feynman's formulation of quantum mechanics as a path-integral [11]. In fact another way to obtain the result is to exploit the transformation (49) to write the Fokker–Planck equation (70) as a Schrödinger equation in imaginary time $\tau = it$, with a potential $U(x) = (1/2)[V'(x)]^2 - \tilde{D}V''(x)$, following Eq. (50). The action in the quantum-mechanical path-integral is

$$\begin{aligned} &\frac{i}{\hbar}\int \mathrm{d}t\left[\frac{1}{2}\dot{x}^2 - U(x)\right] \longrightarrow \\ &\frac{1}{2\tilde{D}}\int \mathrm{d}\tau\left[-\frac{1}{2}\dot{x}^2 - \frac{1}{2}\left[V'(x)\right]^2 + \tilde{D}V''(x)\right]\,, \end{aligned} \tag{84}$$

which is Eq. (82) since $\int_{t_0}^{t}\mathrm{d}t\,\dot{x}V'(x) = \int_{x_0}^{x}\mathrm{d}x\,V'(x) = V(x) - V(x_0)$ does not depend on the path, only on the end-points. The functional $S[x]$ is analogous to the action in classical mechanics, and is frequently referred to as such. It is also sometimes referred to as the generalized Onsager–Machlup functional, in recognition of the original work carried out by Onsager and Machlup [40], in the case of a linear Langevin equation, in 1953.

The above discussion can be generalized in many ways. For example, if the Langevin equation for an n-dimensional process takes the form

$$\dot{x}_i = A_i(\mathbf{x}) + \zeta_i(t)\,, \quad \langle\zeta_i(t)\zeta_j(t')\rangle = 2D_{ij}\delta(t-t')\,, \tag{85}$$

where $\zeta_i(t)$ is a Gaussian noise with zero mean and D_{ij} is independent of $\mathbf{x}$, then the general Onsager–Machlup functional is [21]

$$S[x]=\int \mathrm{d}t\left[\frac{1}{4}\sum_{i,j}\{\dot{x}_i-A_i(\mathbf{x})\}\,D^{-1}_{ij}\,\{\dot{x}_j-A_j(\mathbf{x})\}\right.$$
$$\left.+\frac{1}{2}\sum_i\frac{\partial A_i}{\partial x_i}\right]\,, \qquad (86)$$

if the matrix D_{ij} is non-singular. The generalization to the situation where the noise is multiplicative is more complicated, and is analogous to the path-integral formulation of quantum mechanics in curved space [20].

System Size Expansion

In Example 2 of Sect. "The Fokker–Planck Equation" we explicitly showed how the master equation may have different limits when the size of the system, N, becomes large. In one case both the first and second jump moments were of the same order (and much larger than the higher jump moments) and so a nonlinear Fokker–Planck equation of the diffusion type was obtained in the limit $N \to \infty$. In another case, the first jump moment scaled in a different way to the second moment, and so the $N \to \infty$ limit gave a deterministic macroscopic equation of the form $\dot{x}=f(x)$, with finite N effects presumably consisting of $1/\sqrt{N}$ fluctuations about the macroscopic state, $x(t)$. It is the second scenario that we will explore in this section. It can be formalized by writing

$$\frac{n}{N}=x(t)+\frac{\xi}{\sqrt{N}}\,, \qquad (87)$$

and substituting this into the master equation, then equating terms of the same order in $1/\sqrt{N}$. The leading order equation obtained in this way will be the macroscopic equation, and the function $f(x)$ will emerge from the analysis. The next-to-leading order equation turns out to be a *linear* Fokker–Planck equation in the variable ξ. Higher order terms may also be included. This formalism was first developed by van Kampen [51] and is usually referred to as van Kampen's system-size expansion. We will describe it in the specific case of a one-step process for a single stochastic variable in order to bring out the essential features of the method.

When using this formalism it is useful to rewrite the master equation (23) using step operators which act on an arbitrary function of n according to $\mathcal{E}f(n)=f(n+1)$ and $\mathcal{E}^{-1}f(n)=f(n-1)$. This gives

$$\frac{\mathrm{d}P(n,t)}{\mathrm{d}t}=(\mathcal{E}-1)\left[T(n-1|n)P(n,t)\right]$$
$$+\left(\mathcal{E}^{-1}-1\right)\left[T(n+1|n)P(n,t)\right]\,. \qquad (88)$$

We begin by using Eq. (87) to write the pdf which appears in the master Equation (88) as

$$P(Nx(t)+\sqrt{N}\xi,t)=\Pi(\xi,t)$$
$$\Rightarrow\quad \dot{P}=\frac{\partial\Pi}{\partial t}-N^{1/2}\frac{\mathrm{d}x}{\mathrm{d}t}\frac{\partial\Pi}{\partial\xi}\,. \qquad (89)$$

This gives an expression for the left-hand side of the master equation, $\dot{P}$. To get an expression for the right-hand side,

(a) the step operators are expanded in powers of $1/\sqrt{N}$ [53]:

$$\mathcal{E}^{\pm 1}=1\pm\frac{1}{\sqrt{N}}\frac{\partial}{\partial\xi}+\frac{1}{2!}\frac{1}{N}\frac{\partial^2}{\partial\xi^2}+\mathcal{O}\left(\frac{1}{N^{3/2}}\right)\,, \qquad (90)$$

(b) $T(n\pm 1|n)$ is expressed in terms of ξ and N,
(c) $P(n,t)$ is replaced by $\Pi(\xi,t)$.

Steps (a), (b) and (c) gives the right-hand side of the master equation as a power-series in $1/\sqrt{N}$. Equating the left-hand and right-hand sides order by order in $1/\sqrt{N}$ (this may require a rescaling of the time, t, by a power of $\sqrt{N}$), gives to leading order (the $\partial\Pi/\partial\xi$ cancels) an equation of the form $\mathrm{d}x/\mathrm{d}t=f(x)$. This may be solved subject to the condition $x(0)=x_0=n_0/N$, if we take the initial condition on the master equation to be $P(n,0)=\delta_{n,n_0}$. We denote the solution of this macroscopic equation by $x_{\mathrm{M}}(t)$.

To next order in $1/\sqrt{N}$, the Fokker–Planck equation

$$\frac{\partial\Pi}{\partial t}=-f'(x)\frac{\partial}{\partial\xi}\left[\xi\Pi\right]+\frac{1}{2}g(x)\frac{\partial^2\Pi}{\partial\xi^2}\,, \qquad (91)$$

describing a linear stochastic process is found. The functions $f'(x)$ and $g(x)$ are to be evaluated when $x=x_{\mathrm{M}}(t)$, and so are simply functions of time. If the macroscopic system tends to a fixed point: $x_{\mathrm{M}}(t)\to x^*$, as $t\to\infty$, then $f'(x)$ and $g(x)$ may be replaced by constants in order to study the fluctuations about this stationary state.

To illustrate the method we use Example 3, Sect. "The Master Equation". Equating both sides of the master equation in this case one finds

$$-N^{1/2}\frac{\mathrm{d}x}{\mathrm{d}t}\frac{\partial\Pi}{\partial\xi}+\frac{\partial\Pi}{\partial t}=\frac{1}{\sqrt{N}}\left[f_-(x)-f_+(x)\right]\frac{\partial\Pi}{\partial\xi}$$
$$+\frac{1}{2}\frac{1}{N}\left[f_-(x)+f_+(x)\right]\frac{\partial^2\Pi}{\partial\xi^2}$$
$$+\frac{1}{N}\left[f_-'(x)-f_+'(x)\right]\frac{\partial}{\partial\xi}\left[\xi P\right]+\dots\,, \qquad (92)$$

where the functions $f_-(x)$ and $f_+(x)$ are given by:

$$f_-(x)=cx^2+\mathrm{d}x\,,\qquad f_+(x)=2bx(1-x)\,. \qquad (93)$$

For the left- and right-hand sides to balance in Eq. (92), a rescaled time $\tau = t/N$ needs to be introduced. Then to leading order one finds $\mathrm{d}x/\mathrm{d}\tau = f(x)$, where $f(x) = f_+(x) - f_-(x)$. At next to leading order Eq. (91) is found with $g(x) = f_+(x) + f_-(x)$. The explicit form of the macroscopic equation is

$$\frac{\mathrm{d}x}{\mathrm{d}\tau} = x\,(r - ax)\ , \tag{94}$$

where $r = 2b - d$ and $a = 2b + c$. Equation (94) is the logistic equation, which is the usual phenomenological way to model intraspecies competition.

Since the Fokker–Planck equation (91) describes a linear process, its solution is a Gaussian. This means that the probability distribution $\Pi(\xi, t)$ is completely specified by the first two moments $\langle \xi(t) \rangle$ and $\langle \xi^2(t) \rangle$. Multiplying Eq. (91) by ξ and ξ^2 and integrating over all ξ one finds

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\langle \xi(t)\rangle &= f'(x_{\mathrm{M}}(t))\langle \xi(t)\rangle\ , \\ \frac{\mathrm{d}}{\mathrm{d}t}\langle \xi^2(t)\rangle &= 2f'(x_{\mathrm{M}}(t))\langle \xi^2(t)\rangle + g(x_{\mathrm{M}}(t))\ . \end{aligned} \tag{95}$$

We have chosen the initial condition to be $x_0 = n_0/N$, which implies that $\xi(0) = 0$. The first equation in (95) then implies that $\langle \xi(t) \rangle = 0$ for all t. Multiplying the second equation by $f^{-2}(x_{\mathrm{M}}(t))$ one finds that

$$\langle \xi^2(t)\rangle = f^2(x_{\mathrm{M}}(t)) \int_0^t \mathrm{d}t'\, \frac{g(x_{\mathrm{M}}(t'))}{f^2(x_{\mathrm{M}}(t'))}\ , \tag{96}$$

and so the determination of $\langle \xi^2(t) \rangle$ is reduced to quadrature. The method can be applied to systems with more than one stochastic variable and those which are not one-step processes. Details are given in van Kampen's book [53].

Future Directions

This article has focused largely on classical topics in the theory of stochastic processes, since these form the foundations on which the subject is built. Much of the current work, and one would expect future work, will be numerical in character. Some of this will begin from a basic Markovian description in the form of reactions among chemical species – even if the system is not chemical in nature (Example 3 of Sect. "The Master Equation" is an example). A straightforward algorithm developed by Gillespie [17,18], and since then extended and improved [3,19], provides an efficient way of simulating such systems. It thus provides a valuable method of investigating systems which may be formulated as complicated multivariable master equations, which complements the methods we have discussed here. However, many current studies do not begin from a system which can be described in this way, and there is every indication that this will be more true in the future. For instance, in agent based models the stochastic element may be due to mutations in characteristics, traits or behavior, which may be difficult or impossible to formulate mathematically. Such agent based models are certainly individually based, but each individual may have different attributes and generally behave in such a complex way that only numerical simulations can be used to explore the behavior of the system as a whole. Although these complex systems may be used to model more realistic situations, the well-known problems associated with the large number of parameters typically required to describe such systems, will mean that simplified versions will need to be analyzed in order to understand them at a deeper level. These simpler models are likely to include those where the agents of a particular species are essentially identical. In this article we have discussed how the classical equations of the theory of stochastic processes, such as the Fokker–Planck equation, can be obtained from such models. They will therefore form a bridge between the agent-based approaches which are expected to become more prevalent in the future, and the analytic approaches which lie at the heart of the theory of stochastic processes.

Bibliography

Primary Literature

1. Abramowitz M, Stegun I (Eds) (1965) Handbook of mathematical functions. Dover, New York
2. Bachelier L (1900) Théorie de la spéculation. Annales Scientifiques de L'Ecole Normale Supérieure III(17):21–86
3. Cao Y, Li H, Petzold L (2004) Efficient formulation of the stochastic simulation algorithm for chemically reacting systems. J Chem Phys 121:4059–4067
4. Chandrasekhar S (1943) Stochastic problems in physics and astronomy. Rev Mod Phys 15:1–89. Reprinted in Wax (1954)
5. Cox DR, Miller HD (1968) The theory of stochastic processes, Chap 3. Chapman and Hall, London
6. Crow JF, Kimura M (1970) An introduction to population genetics theory. Harper and Row, New York
7. Ehrenfest P, Ehrenfest T (1907) Über zwei bekannte Einwände gegen das Boltzmannsche H-Theorem. Phys Z 8:311–314
8. Einstein A (1905) Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. Ann Physik 17:549–560. For a translation see: A. Einstein, "Investigations on the Theory of the Brownian Movement" Fürth R (ed), Cowper AD (tr) (Dover, New York, 1956). Chapter I

9. Einstein A (1906) Zur Theorie der Brownschen Bewegung. Ann Physik 19:371–381. For a translation see: A. Einstein, "Investigations on the Theory of the Brownian Movement" Fürth R (ed), Cowper AD (tr) (Dover, New York, 1956). Chapter II
10. Feller W (1968) An introduction to probability theory and its applications, Chap XV, 3rd edn. Wiley, New York
11. Feynman RP (1948) Space-time approach to non-relativistic quantum mechanics. Rev Mod Phys 20:367–387
12. Feynman RP, Hibbs AR (1965) Quantum mechanics and path integrals, Chap 12. McGraw-Hill, New York
13. Fisher RA (1930) The genetical theory of natural selection. Clarendon Press, Oxford
14. Fokker AD (1914) Die mittlere Energie rotierende elektrischer Dipole im Strahlungsfeld. Ann Physik 43:810–820
15. Gantmacher FR (1959) The theory of matrices, Chap 13, Sect 6, vol 2. Chelsea Publishing Co., New York
16. Gardiner CW (2004) Handbook of stochastic methods, 3rd edn. Springer, Berlin
17. Gillespie DT (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J Comput Phys 22:403–434
18. Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. J Phys Chem 81:2340–2361
19. Gillespie DT (2001) Approximate accelerated stochastic simulation of chemically reacting systems. J Chem Phys 115:1716–1733
20. Graham R (1977) Path integral formulation of general diffusion processes. Z Physik B26:281–290
21. Graham R (1975) Macroscopic theory of fluctuations and instabilities. In: Riste T (ed) Fluctuations, Instabilities, and Phase Transitions. Plenum, New York, pp 215–293
22. Haken H (1983) Synergetics. Springer, Berlin. Sect 6.3
23. Kac M (1947) Random walk and the theory of Brownian motion. Amer Math Mon 54:369–391
24. Karlin S, McGregor J (1964) On some stochastic models in genetics. In: Gurland J (ed) Stochastic problems in medicine and biology. University of Wisconsin Press, Madison, pp 245–279
25. Kendall DG (1948) On some modes of population growth leading to R. A. Fisher's logarithmic series distribution. Biometrika 35:6–15
26. Kolmogorov AN (1931) Über die analytischen Methoden in der Wahrscheinlichkeitsrechung. Math Ann 104:415–458
27. Kolmogorov AN (1936) Anfangsgründe der Theorie der Markoffschen Ketten mit unendlich vielen möglichen Zuständen. Mat Sbornik (N.S.) 1:607–610
28. Krafft O, Schaefer M (1993) Mean passage times for tridiagonal transition matrices and a two-parameter Ehrenfest urn model. J Appl Prob 30:964–970
29. Kramers HA (1940) Brownian motion in a field of force and the diffusion model of chemical reactions. Physica 7:284–304
30. Langevin P (1908) Sur la théorie du mouvement brownien. C R Acad Sci Paris 146:530–533. For a translation see: D. S. Lemons and A. Gythiel, Am. J. Phys. 65: 1079–1081 (1997)
31. Malécot G (1944) Sur un problème de probabilités en chaine que pose la génétique. C R Acad Sci Paris 219:379–381
32. Markov AA (1906) Rasprostranenie zakona bol'shih chisel na velichiny, zavisyaschie drug ot druga. Izv Fiz-Matem Obsch Kazan Univ (Series 2) 15:135–156. See also: "Extension of the limit theorems of probability theory to a sum of variables connected in a chain", in Appendix B of R. Howard "Dynamic Probabilistic Systems, vol 1: Markov Chains" (John Wiley and Sons, 1971)
33. May RM (1973) Model ecosystems, Chap 5. Princeton University Press, Princeton
34. McKane AJ, Newman TJ (2004) Stochastic models in population biology and their deterministic analogs. Phys Rev E 70:041902
35. McKane AJ, Newman TJ (2005) Predator-prey cycles from resonant amplification of demographic stochasticity. Phys Rev Lett 94:218102
36. Moran PAP (1958) Random processes in genetics. Proc Cambridge Philos Soc 54:60–72
37. Moran PAP (1962) The statistical processes of evolutionary theory, Chap 4. Clarendon Press, Oxford
38. Moyal JE (1949) Stochastic processes and statistical physics. J Roy Stat Soc (London) B 11:150–210
39. Nordsieck A, Lamb WE Jr, Uhlenbeck GE (1940) On the theory of cosmic-ray showers. I. The Furry model and the fluctuation problem. Physica 7:344–360
40. Onsager L, Machlup S (1953) Fluctuations and irreversible processes. Phys Rev 91:1505–1512
41. Pauli W (1928) Probleme der modernen Physik. In: Debye P (ed) Festschrift zum 60. Geburtstag A. Sommerfeld. Hirzel, Leipzig, p 30
42. Pearson K (1905) The problem of the random walk. Nature 72:294,342
43. Planck M (1917) Über einen Satz der statistischen Dynamik und seine Erweiterung in der Quantentheorie. Abh Preuss Akad Wiss Berl 24:324–341
44. Rayleigh L (1891) Dynamical problems in illustration of the theory of gases. Phil Mag 32:424–445
45. Reichl LE (1998) A modern course in statistical physics, Chap 5, 2nd edn. Wiley, New York
46. Risken H (1989) The Fokker–Planck equation, 2nd edn. Springer, Berlin
47. Schiff LI (1968) Quantum mechanics, Chap 4, 3rd edn. McGraw-Hill, Tokyo
48. Schulman LS (1981) Techniques and applications of path–integration, Chap 5. Wiley, New York
49. Siegert AJF (1949) On the approach to statistical equilibrium. Phys Rev 76:1708–1714
50. Sneddon IN (1957) Elements of partial differential equations, Chap 2. McGraw-Hill, New York
51. van Kampen NG (1961) A power series expansion of the master equation. Can J Phys 39:551–567
52. van Kampen NG (1981) Itô versus Stratonovich. J Stat Phys 24:175–187
53. van Kampen NG (1992) Stochastic processes in physics and chemistry, 2nd edn. North-Holland, Amsterdam
54. von Smoluchowski M (1906) Zur kinetishen Theorie der Brownschen Molekularbewegung und der Suspensionen. Ann Physik 21:756–780
55. von Smoluchowski M (1916) Drei Vortage über Diffusion, Brownsche Bewegung, und Koagulation von Kolloidteilchen. Phys Z 17:571–599
56. Wiener N (1921) The average of an analytic functional. Proc Natl Acad Sci USA 7:253–260
57. Wiener N (1921) The average of an analytic functional and the Brownian movement. Proc Natl Acad Sci USA 7:294–298

58. Wright S (1931) Evolution in Mendelian populations. Genetics 16:97–159
59. Zinn-Justin J (2002) Quantum field theory and critical phenomena, 4th edn. Clarendon Press, Oxford. Sect 4.8.2
60. Zwanzig R (1973) Nonlinear generalized Langevin equations. J Stat Phys 9:215–220

Books and Reviews

Wax N (1954) Selected papers on noise and stochastic processes. Dover, New York

Stochastic Volatility

TORBEN G. ANDERSEN[1,2,3], LUCA BENZONI[4]

[1] Kellogg School of Management, Northwestern University, Evanston, USA
[2] NBER, Cambridge, USA
[3] CREATES, Aarhus, Denmark
[4] Federal Reserve Bank of Chicago, Chicago, USA

Article Outline

Glossary

Implied volatility The value of asset return volatility which equates a model-implied derivative price to the observed market price. Most notably, the term is used to identify the volatility implied by the Black and Scholes [63] option pricing formula.

Quadratic return variation The ex-post sample-path return variation over a fixed time interval.

Realized volatility The sum of finely sampled squared asset return realizations over a fixed time interval. It is an estimate of the quadratic return variation over such time interval.

Stochastic volatility A process in which the return variation dynamics include an unobservable shock which cannot be predicted using current available information.

Definition of the Subject

Given the importance of return volatility on a number of practical financial management decisions, the efforts to provide good real-time estimates and forecasts of current and future volatility have been extensive. The main framework used in this context involves stochastic volatility models. In a broad sense, this model class includes GARCH, but we focus on a narrower set of specifications in which volatility follows its own random process, as is common in models originating within financial economics. The distinguishing feature of these specifications is that volatility, being inherently unobservable and subject to independent random shocks, is not measurable with respect to observable information. In what follows, we refer to these models as *genuine* stochastic volatility models.

Much modern asset pricing theory is built on continuous-time models. The natural concept of volatility within this setting is that of genuine stochastic volatility. For example, stochastic volatility (jump-)diffusions have provided a useful tool for a wide range of applications, including the pricing of options and other derivatives, the modeling of the term structure of risk-free interest rates, and the pricing of foreign currencies and defaultable bonds. The increased use of intraday transaction data for construction of so-called realized volatility measures provides additional impetus for considering genuine stochastic volatility models. As we demonstrate below, the realized volatility approach is closely associated with the continuous-time stochastic volatility framework of financial economics.

There are some unique challenges in dealing with genuine stochastic volatility models. For example, volatility is truly latent and this feature complicates estimation and inference. Further, the presence of an additional state variable – volatility – renders the model less tractable from an analytic perspective. We review how such challenges have been addressed through development of new estimation methods and imposition of model restrictions allowing for closed-form solutions while remaining consistent with the dominant empirical features of the data.

Introduction

The label Stochastic Volatility is applied in two distinct ways in the literature. For one, it is used to signify that the (absolute) size of the innovations of a time series displays random fluctuations over time. Descriptive models of financial time series almost invariably embed this feature nowadays as asset return series tend to display alternating quiet and turbulent periods of varying length and intensity. To distinguish this feature from models that operate with an a priori known or deterministic path for the volatility process, the random evolution of the conditional return variance is termed stochastic volatility. The simplest case of deterministic volatility is the constant variance assumption invoked in, e. g., the Black and Scholes [63] framework. Another example is modeling the variance purely as a given function of calendar time, allowing only for effects such as time-of-year (seasonals), day-of-week (institutional and announcement driven) or time-of-day (diurnal effects due to, e. g., market microstructure features). Any model not falling within this class is then a stochastic volatility model. For example, in the one-factor continuous-time Cox, Ingersoll, and Ross [113] (CIR) model the (stochastic) level of the short term interest rate governs the dynamics of the (instantaneous) drift and diffusion term of all zero-coupon yields. Likewise, in GARCH models the past return innovations govern the one-period ahead conditional mean and variance. In both models, the volatility is known, or deterministic, at a given point in time, but the random evolution of the processes renders volatility stochastic for any horizon beyond the present period.

The second notion of stochastic volatility, which we adopt henceforth, refers to models in which the return variation dynamics is subject to an unobserved random shock so that the volatility is inherently latent. That is, the current volatility state is not known for sure, conditional on the true data generating process and the past history of all available discretely sampled data. Since the CIR and GARCH models described above render the current (conditional) volatility known, they are not stochastic volatility models in this sense. In order to make the distinction clear cut, we follow Andersen [10] and label this second, more restrictive, set *genuine* stochastic volatility (SV) models.

There are two main advantages to focusing on SV models. First, much asset pricing theory is built on continuous-time models. Within this class, SV models tend to fit more naturally with a wide array of applications, including the pricing of currencies, options, and other derivatives, as well as the modeling of the term structure of interest rates. Second, the increasing use of high-frequency intraday data for construction of so-called realized volatility measures is also starting to push the GARCH models out of the limelight as the realized volatility approach is naturally linked to the continuous-time SV framework of financial economics.

One drawback is that volatility is not measurable with respect to observable (past) information in the SV setting. As such, an estimate of the current volatility state must be filtered out from a noisy environment and the estimate will change as future observations become available. Hence, in-sample estimation typically involves smoothing techniques, not just filtering. In contrast, the conditional variance in GARCH is observable given past information, which renders (quasi-)maximum likelihood techniques for inference quite straightforward while smoothing techniques have no role. As such, GARCH models are easier to estimate and practitioners often rely on them for time-series forecasts of volatility. However, the development of powerful method of simulated moments, Markov Chain Monte Carlo (MCMC) and other simulation based procedures for estimation and forecasting of SV models may well render them competitive with ARCH over time on that dimension.

Direct indications of the relations between SV and GARCH models are evident in the sequence of papers by Dan Nelson and Dean Foster exploring the SV diffusion limits of ARCH models as the case of continuous sampling is approached, see, e. g., Nelson and Foster [219]. Moreover, as explained in further detail in the estimation section below, it can be useful to summarize the dynamic features of asset returns by tractable pseudo-likelihood scores obtained from GARCH-style models when performing simulation based inference for SV models. As such, the SV and GARCH frameworks are closely related and should be viewed as complements. Despite these connections we focus, for the sake of brevity, almost exclusively on SV models and refer the interested reader to the GARCH chapter for further information.

The literature on SV models is vast and rapidly growing, and excellent surveys are available, e. g., Ghysels et al. [158] and Shephard [239,240]. Consequently, we focus on providing an overview of the main approaches with illustrations of the scope for applications of these models to practical finance problems.

Model Specification

The original econometric studies of SV models were invariably cast in discrete time and they were quite similar in structure to ARCH models, although endowed with a more explicit structural interpretation. Recent work in the area has been mostly directly towards a continuous time setting and motivated by the typical specifications in financial economics. This section briefly reviews the two alternative approaches to specification of SV models.

Discrete-Time SV Models and the Mixture-of-Distributions Hypothesis

Asset pricing theory contends that financial asset prices reflect the discounted value of future expected cash flows, implying that all news relevant for either discount rates or cash flows should induce a shift in market prices. Since economic news items appear almost continuously in real time, this perspective rationalizes the ever-changing nature of prices observed in financial markets. The process linking news arrivals to price changes may be complex, but if it is stationary in the statistical sense it will nonetheless produce a robust theoretical association between news arrivals, market activity and return volatility. In fact, if the number of news arrival is very large, standard central limit theory will tend to imply that asset returns are approximately normally distributed *conditional* on the news count. More generally, variables such as the trading volume, the number of transactions or the number of price quotes are also naturally related to the intensity of the information flow. This line of reasoning has motivated specifications such as

$$y_t|s_t \rightsquigarrow N(\mu_y s_t \ , \ \sigma_y^2 s_t) \ , \tag{1}$$

where y_t is an "activity" variable related to the information flow, s_t is a positive intensity process reflecting the rate of news arrivals, μ_y represents the mean response to an information event, and σ_y is a pure scaling parameter.

This is a normal mixture model, where the s_t process governs or "mixes" the scale of the distribution across the periods. If s_t is constant, this is simply an i.i.d. Gaussian process for returns and possible other related variables. However, this is clearly at odds with the empirical evidence for, e. g., return volatility and trading volume. Therefore, s_t is typically stipulated to follow a separate stochastic process with random innovations. Hence, each period the return series is subject to two separate shocks, namely the usual idiosyncratic error term associated with the (normal) return distribution, but also a shock to the variance or volatility process, s_t. This endows the return process with genuine stochastic volatility, reflecting the random intensity of news arrivals. Moreover, it is typically assumed that only returns, transactions and quotes are observable, but not the actual value of s_t itself, implying that σ_y cannot be separately identified. Hence, we simply fix this parameter at unity.

The time variation in the information flow series induces a fat-tailed unconditional distribution, consistent with stylized facts for financial return and, e. g., trading volume series. Intuitively, days with a lot of news display more rapid price fluctuations and trading activity than days with a low news count. In addition, if the s_t process is positively correlated, then shocks to the conditional mean and variance processes for y_t will be persistent. This is consistent with the observed clustering in financial markets, where return volatility and trading activity are contemporaneously correlated and each display pronounced positive serial dependence.

The inherent randomness and unobserved nature of the news arrival process, even during period t, renders the true mean and variance series latent. This property is the major difference with the GARCH model class, in which the one-step-ahead conditional mean and variance are a known function of observed variables at time $t-1$. As such, for genuine SV models, we must distinguish the full, but infeasible, information set ($s_t \in \mathcal{F}_t$) and the observable information set ($s_t \notin \mathcal{I}_t$). This basic latency of the mixing variable (state vector) of the SV model complicates inference and forecasting procedures as discussed below.

For short horizon returns, μ_y is nearly negligible and can reasonably be ignored or simply fixed at a small constant value, and the series can then be demeaned. This simplification produces the following return (innovation) model,

$$r_t = \sqrt{s_t}\, z_t \ , \tag{2}$$

where z_t is an i.i.d. standard normal variable, implying a simple normal-mixture representation,

$$r_t|s_t \rightsquigarrow N(0, s_t) \ . \tag{3}$$

Univariate return models of the form (3) as well as multivariate systems including a return variable along with other related market activity variables, such as the transactions count, the quote intensity or the aggregate trading volume, stem from the Mixture-of-Distributions Hypothesis (MDH).

Actual implementation of the MDH hinges on a particular representation of the information-arrival process s_t. Clark [102] uses trading volume as a proxy for the activity variable, a choice motivated by the high contemporaneous correlation between return volatility and volume. Tauchen and Pitts [247] follow a structural approach to characterize the joint distribution of the daily return and volume relation governed by the underlying latent information flow s_t. However, both these models assume temporal independence of the information flow, thus failing to capture the clustering in these series. Partly in response, Gallant et al. [153] examine the joint conditional return-volume distribution without imposing any structural MDH restrictions. Nonetheless, many of the original discrete-time SV specifications are compatible with the MDH framework, including Taylor [249][1], who proposes an autoregressive parametrization of the latent log-volatility (or information flow) variable

$$\log(s_{t+1}) = \eta_0 + \eta_1 \log(s_t) + u_t,\ u_t \rightsquigarrow \text{i.i.d}(0, \sigma_u^2)\,, \tag{4}$$

[1] Discrete-time SV models go father back in time, at least to the easly paper by Rosenberg [232] recently reprinted in Shephard [240].

where the error term, u_t, may be correlated with the disturbance term, z_t, in the return Eq. (2) so that $\rho = \text{corr}(u_t, z_t) \neq 0$. If $\rho < 0$, downward movements in asset prices result in higher future volatility as also predicted by the so-called 'leverage effect' in the exponential GARCH, or EGARCH, form of Nelson [218] and the asymmetric GARCH model of Glosten et al. [160].

Early tests of the MDH include Lamoureux and Lastrapes [194] and Richardson and Smith [231]. Subsequently, Andersen [11] studies a modified version of the MDH that provides a much improved fit to the data. Further refinements of the MDH specification have been pursued by, e. g., Liesenfeld [198,199] and Bollerslev and Jubinsky [67]. Among the first empirical studies of the related approach of stochastic time changes are Ané and Geman [29], who focus on stock returns, and Conley et al. [109], who focus on the short-term risk-free interest rate.

Continuous-Time Stochastic Volatility Models

Asset returns typically contain a predictable component, which compensates the investor for the risk of holding the security, and an unobservable shock term, which cannot be predicted using current available information. The conditional asset return variance pertains to the variability of the unobservable shock term. As such, over a non-infinitesimal horizon it is necessary to first specify the conditional mean return (e. g., through an asset pricing model) in order to identify the conditional return variation. In contrast, over an infinitesimal time interval this is not necessary because the requirement that market prices do not admit arbitrage opportunities implies that return innovations are an order of magnitude larger than the mean return. This result has important implications for the approach we use to model and measure volatility in continuous time.

Consider an asset with log-price process $\{p(t) \ , \ t \in [0, T]\}$ defined on a probability space $(\Omega, \mathcal{F}, P)$. Following Andersen et al. [19] we define the continuously compounded asset return over a time interval from $t-h$ to t, $0 \leq h \leq t \leq T$, to be

$$r(t,h) = p(t) - p(t-h) \,. \tag{5}$$

A special case of (5) is the cumulative return up to time t, which we denote $r(t) \equiv r(t,t) = p(t) - p(0)$, $0 \leq t \leq T$. Assume the asset trades in a frictionless market void of arbitrage opportunities and the number of potential discontinuities (jumps) in the price process per unit time is finite. Then the log-price process p is a semi-martingale (e. g., Back [33]) and therefore the cumulative return $r(t)$ admits the decomposition (e. g., Protter [229])

$$r(t) = \mu(t) + M^{\mathrm{C}}(t) + M^{\mathrm{J}}(t) \,, \tag{6}$$

where $\mu(t)$ is a predictable and finite variation process, $M^{\mathrm{C}}(t)$ a continuous-path infinite-variation martingale, and $M^{\mathrm{J}}(t)$ is a compensated finite activity jump martingale. Over a discrete time interval the decomposition (6) becomes

$$r(t,h) = \mu(t,h) + M^{\mathrm{C}}(t,h) + M^{\mathrm{J}}(t,h) \,, \tag{7}$$

where $\mu(t,h) = \mu(t) - \mu(t-h)$, $M^{\mathrm{C}}(t,h) = M^{\mathrm{C}}(t) - M^{\mathrm{C}}(t-h)$, and $M^{\mathrm{J}}(t,h) = M^{\mathrm{J}}(t) - M^{\mathrm{J}}(t-h)$.

Denote now with $[r,r]$ the quadratic variation of the semi-martingale process r, where (Protter [229])

$$[r,r]_t = r(t)^2 - 2\int r(s-)\mathrm{d}r(s) \,, \tag{8}$$

and $r(t-) = \lim_{s \uparrow t} r(s)$. If the finite variation process μ is continuous, then its quadratic variation is identically zero and the predictable component μ in decomposition (7) does not affect the quadratic variation of the return r. Thus, we obtain an expression for the quadratic return variation over the time interval from $t-h$ to t, $0 \leq h \leq t \leq T$ (e. g., Andersen et al. [21] and Barndorff-Nielsen and Shephard [51,52]):

$$\begin{aligned} \mathrm{QV}(t,h) &= [r,r]_t - [r,r]_{t-h} \\ &= [M^{\mathrm{C}}, M^{\mathrm{C}}]_t - [M^{\mathrm{C}}, M^{\mathrm{C}}]_{t-h} \\ &\quad + \sum_{t-h<s\leq t} \Delta M^2(s) \\ &= [M^{\mathrm{C}}, M^{\mathrm{C}}]_t - [M^{\mathrm{C}}, M^{\mathrm{C}}]_{t-h} \\ &\quad + \sum_{t-h<s\leq t} \Delta r^2(s) \,. \end{aligned} \tag{9}$$

Most continuous-time models for asset returns can be cast within the general setting of Eq. (7), and Eq. (9) provides a framework to study the model-implied return variance. For instance, the Black and Scholes [63] model is a special case of the setting described by Eq. (7) in which the conditional mean process μ is constant, the continuous martingale M^C is a standard Brownian motion process, and the jump martingale M^J is identically zero:

$$\mathrm{d}p(t) = \mu \mathrm{d}t + \sigma \mathrm{d}W(t) \,. \tag{10}$$

In this case, the quadratic return variation over the time interval from $t-h$ to t, $0 \leq h \leq t \leq T$, simplifies to

$$\mathrm{QV}(t,h) = \int_{t-h}^{t} \sigma^2 \mathrm{d}s = \sigma^2 h \,, \tag{11}$$

that is, return volatility is constant over any time interval of length h.

A second notable example is the jump-diffusion model of Merton [214],

$$\mathrm{d}p(t) = (\mu - \lambda\overline{\xi})\mathrm{d}t + \sigma \mathrm{d}W(t) + \xi(t)\mathrm{d}q_t \,, \tag{12}$$

where q is a Poisson process uncorrelated with W and governed by the constant jump intensity λ, i. e., $\mathrm{Prob}(\mathrm{d}q_t = 1) = \lambda \mathrm{d}t$. The scaling factor $\xi(t)$ denotes the magnitude of the jump in the return process if a jump occurs at time t. It is assumed to be normally distributed,

$$\xi(t) \rightsquigarrow N(\overline{\xi}, \sigma_{\xi}^2) \,. \tag{13}$$

In this case, the quadratic return variation process over the time interval from $t-h$ to t, $0 \le h \le t \le T$ becomes

$$\begin{aligned} \mathrm{QV}(t,h) &= \int_{t-h}^{t} \sigma^2 \mathrm{d}s + \sum_{t-h \le s \le t} J(s)^2 \\ &= \sigma^2 h + \sum_{t-h \le s \le t} J(s)^2 \,, \end{aligned} \tag{14}$$

where $J(t) \equiv \xi(t)\mathrm{d}q(t)$ is non-zero only if a jump actually occurs.

Finally, a broad class of stochastic volatility models is defined by

$$\mathrm{d}p(t) = \mu(t)\mathrm{d}t + \sigma(t)\mathrm{d}W(t) + \xi(t)\mathrm{d}q_t \,, \tag{15}$$

where q is a constant-intensity Poisson process with log-normal jump amplitude (13). Equation (15) is also a special case of (7) and the associated quadratic return variation over the time interval from $t-h$ to t, $0 \le h \le t \le T$, is

$$\begin{aligned} \mathrm{QV}(t,h) &= \int_{t-h}^{t} \sigma(s)^2 \mathrm{d}s + \sum_{t-h \le s \le t} J(s)^2 \\ &\equiv IV(t,h) + \sum_{t-h \le s \le t} J(s)^2 \,. \end{aligned} \tag{16}$$

As in the general case of Eq. (9), Eq. (16) identifies the contribution of diffusive volatility, termed 'integrated variance' (IV), and cumulative squared jumps to the total quadratic variation.

Early applications typically ignored jumps and focused exclusively on the integrated variance component. For instance, IV plays a key role in Hull and White's [174] SV option pricing model, which we discuss in Sect. "Options" below along with other option pricing applications. For illustration, we focus here on the SV model specification by Wiggins [256]:

$$\mathrm{d}p(t) = \mu \mathrm{d}t + \sigma(t)\mathrm{d}W_p(t) \tag{17}$$

$$\mathrm{d}\sigma(t) = f(\sigma(t))\mathrm{d}t + \eta\sigma(t)\mathrm{d}W_\sigma(t) \,, \tag{18}$$

where the innovations to the return $\mathrm{d}p$ and volatility σ, W_p and W_σ, are standard Brownian motions. If we define $y = \log(\sigma)$ and apply Itô's formula we obtain

$$\begin{aligned} \mathrm{d}y(t) &= \mathrm{d}\log(\sigma(t)) \\ &= \left[-\frac{1}{2}\eta^2 + \frac{f(\sigma(t))}{\sigma(t)} \right] \mathrm{d}t + \eta \mathrm{d}W_\sigma(t) \,. \end{aligned} \tag{19}$$

Wiggins approximates the drift term $f(\sigma(t)) \approx \{\alpha + \kappa[\log(\overline{\sigma}) - \log(\sigma(t))]\}\sigma(t)$. Substitution in Eq. (19) yields

$$\mathrm{d}\log(\sigma(t)) = [\overline{\alpha} - \kappa \log(\sigma(t))]\mathrm{d}t + \eta \mathrm{d}W_\sigma(t) \,, \tag{20}$$

where $\overline{\alpha} = \alpha + \kappa\log(\overline{\sigma}) - \frac{1}{2}\eta^2$. As such, the logarithmic standard deviation process in Wiggins has diffusion dynamics similar in spirit to Taylor's discrete time AR(1) model for the logarithmic information process, Eq. (4). As in Taylor's model, negative correlation between return and volatility innovations, $\rho = \mathrm{corr}(W_p, W_\sigma) < 0$, generates an asymmetric response of volatility to return shocks similar to the leverage effect in discrete-time EGARCH models.

More recently, several authors have imposed restrictions on the continuous-time SV jump-diffusion (15) that render the model more tractable while remaining consistent with the empirical features of the data. We return to these models in Sect. "Options" below.

Realized Volatility

Model-free measures of return variation constructed only from concurrent return realizations have been considered at least since Merton [215]. French et al. [148] construct monthly historical volatility estimates from daily return observations. More recently, the increased availability of transaction data has made it possible to refine early measures of historical volatility into the notion of 'realized volatility', which is endowed with a formal theoretical justification as an estimator of the quadratic return variation as first noted in Andersen and Bollerslev [18]. The realized volatility of an asset return r over the time interval from $t-h$ to t is

$$RV(t,h;n) = \sum_{i=1}^{n} r\left(t - h + \frac{ih}{n}, \frac{h}{n}\right)^2 \,. \tag{21}$$

Semi-martingale theory ensures that the realized volatility measure RV converges to the return quadratic variation QV, previously defined in Eq. (9), when the sampling frequency n increases. We point the interested reader to, e. g., Andersen et al. [19] to find formal arguments in support of this claim. Here we convey intuition for this result by considering the special case in which the asset return follows a continuous-time diffusion without jumps,

$$\mathrm{d}p(t) = \mu(t)\mathrm{d}t + \sigma(t)\mathrm{d}W(t)\,. \tag{22}$$

As in Eq. (21), consider a partition of the $[t-h, t]$ interval with mesh h/n. A discretization of the diffusion (22) over a sub-interval from $(t-h+(i-1)h/n)$ to $(t-h+ih/n)\,,\ i = 1,\dots,n$, yields

$$\begin{aligned} r\left(t-h+\frac{ih}{n},\frac{h}{n}\right) &\approx \mu\left(t-h+\frac{(i-1)h}{n}\right)\frac{h}{n} \\ &+ \sigma\left(t-h+\frac{(i-1)h}{n}\right)\Delta W\left(t-h+\frac{ih}{n}\right)\,, \end{aligned} \tag{23}$$

where $\Delta W(t-h+ih/n) = W(t-h+ih/n) - W(t-h+(i-1)h/n)$.

Suppressing time indices, the squared return r^2 over the time interval of length h/n is therefore:

$$r^2 = \mu^2\left(\frac{h}{n}\right)^2 + 2\mu\sigma\Delta W\left(\frac{h}{n}\right) + \sigma^2(\Delta W)^2\,. \tag{24}$$

As $n \to \infty$ the first two terms vanish at a rate higher than the last one. In particular, to a first order approximation the squared return equals the squared return innovation and therefore the squared return conditional mean and variance are

$$\mathrm{E}\left[r^2|\mathcal{F}_t\right] \approx \sigma^2\frac{h}{n} \tag{25}$$

$$\mathrm{Var}\left[r^2|\mathcal{F}_t\right] \approx 2\sigma^4\left(\frac{h}{n}\right)^2\,. \tag{26}$$

The no-arbitrage condition implies that return innovations are serially uncorrelated. Thus, summing over $i = 1,\dots,n$ we obtain

$$\begin{aligned} &\mathrm{E}\left[RV(t,h,n)|\mathcal{F}_t\right] \\ &\quad = \sum_{i=1}^{n}\mathrm{E}\left[r\left(t-h+\frac{ih}{n},\frac{h}{n}\right)^2|\mathcal{F}_t\right] \\ &\quad \approx \sum_{i=1}^{n}\sigma\left(t-h+\frac{(i-1)h}{n}\right)^2\frac{h}{n} \\ &\quad \approx \int_{t-h}^{t}\sigma(s)^2\mathrm{d}s \end{aligned} \tag{27}$$

$$\begin{aligned} &\mathrm{Var}\left[RV(t,h,n)|\mathcal{F}_t\right] \\ &\quad = \sum_{i=1}^{n}\mathrm{Var}\left[r\left(t-h+\frac{ih}{n},\frac{h}{n}\right)^2|\mathcal{F}_t\right] \\ &\quad \approx \sum_{i=1}^{n}2\sigma\left(t-h+\frac{(i-1)h}{n}\right)^4\left(\frac{h}{n}\right)^2 \\ &\quad \approx 2\left(\frac{h}{n}\right)\int_{t-h}^{t}\sigma(s)^4\mathrm{d}s\,. \end{aligned} \tag{28}$$

Equation (27) illustrates that realized volatility is an unbiased estimator of the return quadratic variation, while Eq. (28) shows that the estimator is consistent as its variance shrinks to zero when we increase the sampling frequency n and keep the time interval h fixed. Taken together, these results suggest that RV is a powerful and model-free measure of the return quadratic variation. Effectively, RV gives practical empirical content to the latent volatility state variable underlying the models previously discussed in Sect. “Continuous-Time Stochastic Volatility Models”.

Two issues complicate the practical application of the convergence results illustrated in Eqs. (27) and (28). First, a continuum of instantaneous return observations must be used for the conditional variance in Eq. (28) to vanish. In practice, only a discrete price record is observed, and thus an inevitable discretization error is present. Barndorff-Nielsen and Shephard [52] develop an asymptotic theory to assess the effect of this error on the RV estimate (see also [209]). Second, market microstructure effects (e. g., price discreteness, bid-ask spread positioning due to dealer inventory control, and bid-ask bounce) contaminate the return observations, especially at the ultra-high frequency. These effects tend to generate spurious correlations in the return series which can be partially eliminated by filtering the data prior to forming the RV estimates. However, this strategy is not a panacea and much current work studies the optimal sampling scheme and the construction of improved realized volatility in the presence of microstructure noise. This growing literature is surveyed by Hansen and Lunde [165], Bandi and Russell [46], McAleer and Medeiros [205], and Andersen and Benzoni [14]. Recent notable contributions to this literature include Bandi and Russell [45], Barndorff-Nielsen et al. [49], Diebold and Strasser [121], and Zhang, Mykland, and Aï t-Sahalia [262]. Related, there is the issue of how to construct RV measures when the market is rather illiquid. One approach is to use a lower sampling frequency and focus on longer-horizon RV measure. Alternatively the literature has explored volatility measures that are more robust to situations in which the noise-to-signal ratio is high, e. g., Alizadeh et al. [8], Brandt and Diebold [72], Brandt and Jones [73], Gallant et al. [151], Garman and Klass [157], Parkinson [221], Schwert [237], and Yang and Zhang [259] consider the high-low price range measure. Dobrev [122] generalizes the range estimator to high-frequency data and shows its link with RV measures.

Equations (27) and (28) also underscore an important difference between RV and other volatility measures. RV is an ex-post model-free estimate of the quadratic variation process. This is in contrast to ex-ante measures which attempt to forecast future quadratic variation using information up to current time. The latter class includes parametric GARCH-type volatility forecasts as well as forecasts built from stochastic volatility models through, e. g., the Kalman filter (e. g., [167,168]), the particle filter (e. g., [186,187]) or the reprojection method (e. g., [152,155]).

More recently, other studies have pursued more direct time-series modeling of volatility to obtain alternative ex-ante forecasts. For instance, Andersen et al. [21] follow an ARMA-style approach, extended to allow for long memory features, to model the logarithmic foreign exchange rate realized volatility. They find the fit to dominate that of traditional GARCH-type models estimated from daily data. In a related development, Andersen, Bollerslev, and Meddahi [24,25] exploit the general class of Eigenfunction Stochastic Volatility (ESV) models introduced by Meddahi [208] to provide optimal analytic forecast formulas for realized volatility as a function of past realized volatility. Other scholars have pursued more general model specifications to improve forecasting performance. Ghysels et al. [159] consider Mixed Data Sampling (MIDAS) regressions that use a combination of volatility measures estimates at different frequencies and horizons. Related, Engle and Gallo [137] exploit the information in different volatility measures, captured by a multivariate extension of the multiplicative error model suggested by Engle [136], to predict multi-step volatility. Finally, Andersen et al. [20] build on the Heterogeneous AutoRegressive (HAR) model by Barndorff-Nielsen and Shephard [50] and Corsi [110] and propose a HAR-RV component-based regression to forecast the h-steps ahead quadratic variation:

$$\begin{aligned}\mathrm{RV}(t+h,h) = \beta_0 + \beta_D \mathrm{RV}(t,1) + \beta_W \mathrm{RV}(t,5)\\ + \beta_M \mathrm{RV}(t,21) + \varepsilon(t+h)\,. \qquad (29)\end{aligned}$$

Here the lagged volatility components $\mathrm{RV}(t,1)$, $\mathrm{RV}(t,5)$, and $\mathrm{RV}(t,21)$ combine to provide a parsimonious approximation to the long-memory type behavior of the realized volatility series, which has been documented in several studies (e. g., Andersen et al. [19]). Simple OLS estimation yields consistent estimates for the coefficients in the regression (29), which can be used to forecast volatility out of sample.

As mentioned previously, the convergence results illustrated in Eqs. (27) and (28) stem from the theory of semi-martingales under conditions more general than those underlying the continuous-time diffusion in Eq. (22). For instance, these results are robust to the presence of discontinuities in the return path as in the jump-diffusion SV model (15). In this case the realized volatility measure (21) still converges to the return quadratic variation, which is now the sum of the diffusive integrated volatility IV and the cumulative squared jump component:

$$\mathrm{QV}(t,h) = \mathrm{IV}(t,h) + \sum_{t-h\le s\le t} J(s)^2\,. \qquad (30)$$

The decomposition in Eq. (30) motivates the quest for separate estimates of the two quadratic variation components, IV and squared jumps. This is a fruitful exercise in forecasting applications, since separate estimation of the two components increases predictive accuracy (e. g., [20]). Further, this decomposition is relevant for derivatives pricing, e. g., options are highly sensitive to jumps as well as large moves in volatility (e. g., [141,220]).

A consistent estimate of integrated volatility is the k-skip bipower variation, BV (e. g., Barndorff-Nielsen and Shephard [53]),

$$\begin{aligned}\mathrm{BV}(t,h;k,n) = \frac{\pi}{2}\sum_{i=k+1}^{n}\left| r\left(t-h+\frac{ih}{n},\frac{h}{n}\right)\right|\\ \times\left| r\left(t-h+\frac{(i-k)h}{n},\frac{h}{n}\right)\right|. \qquad (31)\end{aligned}$$

Liu and Maheu [202] and Forsberg and Ghysels [147] show that realized power variation, which is robust to the presence of jumps, can improve volatility forecasts. A well-known special case of (31) is the ‘realized bipower variation’, which has $k = 1$ and is denoted $\mathrm{BV}(t,h;n) \equiv \mathrm{BV}(t,h;1,n)$. We can combine bipower variation with the realized volatility RV to obtain a consistent estimate of the squared jump component, i. e.,

$$\begin{aligned}\mathrm{RV}(t,h;n) - \mathrm{BV}(t,h;n) \underset{n\to\infty}{\longrightarrow} \mathrm{QV}(t,h) - \mathrm{IV}(t,h)\\ = \sum_{t-h\le s\le t} J(s)^2\,. \qquad (32)\end{aligned}$$

The result in Eq. (32) are useful to design tests for the presence of jumps in volatility, e. g., Andersen et al. [20], Barndorff-Nielsen and Shephard [53,54], Huang and

Tauchen [172], and Mizrach [217]. More recently, alternative approaches to test for jumps have been developed by Aït-Sahalia and Jacod [6], Andersen et al. [23], Lee and Mykland [195], and Zhang [261].

Applications

The power of the continuous-time paradigm has been evident ever since the work by Merton [212] on intertemporal portfolio choice, Black and Scholes [63] on option pricing, and Vasicek [255] on bond valuation. However, the idea of casting these problems in a continuous-time diffusion context goes back all the way to the work in 1900 by Bachelier [32].

Merton [213] develops a continuous-time general-equilibrium intertemporal asset pricing model which is later extended by Cox et al. [112] to a production economy. Because of its flexibility and analytical tractability, the Cox et al. [112] framework has become a key tool used in several financial applications, including the valuation of options and other derivative securities, the modeling of the term structure of risk-free interest rates, the pricing of foreign currencies and defaultable bonds.

Volatility has played a central role in these applications. For instance, an option's payoff is non-linear in the price of the underlying asset and this feature renders the option value highly sensitive to the volatility of underlying returns. Further, derivatives markets have grown rapidly in size and complexity and financial institutions have been facing the challenge to manage intricate portfolios exposed to multiple risk sources. Risk management of these sophisticated positions hinges on volatility modeling. More recently, the markets have responded to the increasing hedging demands of investors by offering a menu of new products including, e. g., volatility swaps and derivatives on implied volatility indices like the VIX. These innovations have spurred an even more pressing need to accurately measure and forecast volatility in financial markets.

Research has responded to these market developments. We next provide a brief illustrative overview of the recent literature dealing with option pricing and term structure modeling, with an emphasis on the role that volatility modeling has played in these two key applications.

Options

Rubinstein [233] and Bates [55], among others, note that prior to the 1987 market crash the Black and Scholes [63] (BS) formula priced option contracts quite accurately whereas after the crash it has been systematically underpricing out-of-the-money equity-index put contracts. This feature is evident from Fig. 1, which is constructed from options on the S&P 500 futures. It shows the implied volatility function for near-maturity contracts traded both before and after October 19, 1987 ('Black Monday'). The mild u-shaped pattern prevailing in the pre-crash implied volatilities is labeled a 'volatility smile,' in contrast to the asymmetric post-1987 'volatility smirk'. Importantly, while the steepness and level of the implied volatility curve fluctuate day to day depending on market conditions, the curve has been asymmetric and upward sloping ever since 1987, so the smirk remains in place to the current date, e. g., Benzoni et al. [60]. In contrast, before the crash the implied volatility curve was invariably flat or mildly u-shaped as documented in, e. g., [57]. Finally, we note that the post-1987 asymmetric smirk for index options contrasts sharply with the pattern for individual equity options, which possess flat or mildly u-shaped implied volatility curves (e. g., [37,65]).

Given the failures of the BS formula, much research has gone into relaxing the underlying assumptions. A natural starting point is to allow volatility to evolve randomly, inspiring numerous studies that examine the option pricing implications of SV models. The list of early contributions includes [174,188,211,238,244,245,256]. Here we focus in particular on the Hull and White [174] model,

$$\mathrm{d}p(t) = \mu_p \mathrm{d}t + \sqrt{V(t)}\mathrm{d}W_p(t) \tag{33}$$

$$\frac{\mathrm{d}V(t)}{V(t)} = \mu_V \mathrm{d}t + \sigma_V \mathrm{d}W_V(t) \,, \tag{34}$$

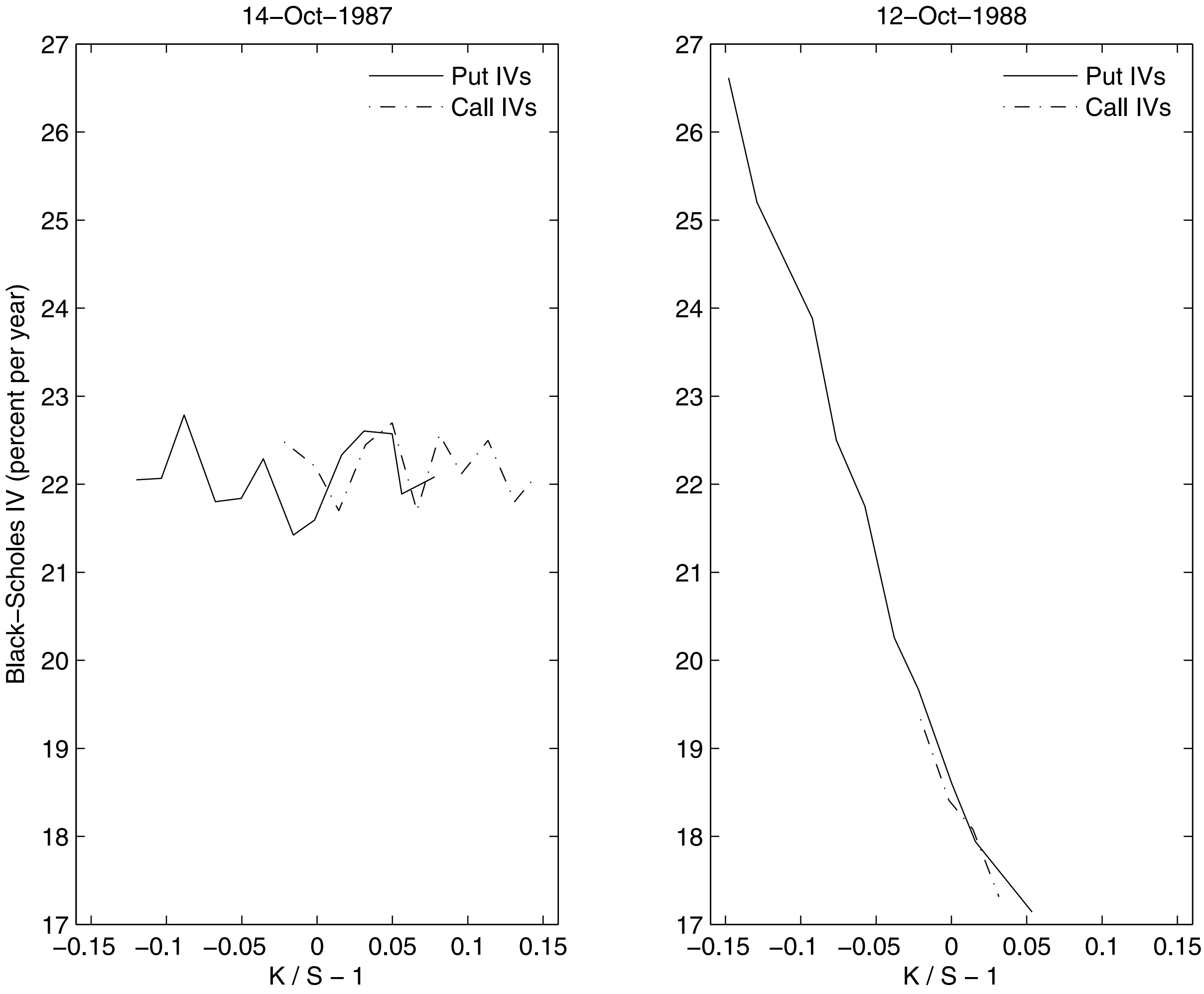
where W_p and W_V are standard Brownian motions. In general, shocks to returns and volatility may be (negatively) correlated, however for tractability Hull and White assume $\rho = \mathrm{corr}(\mathrm{d}W_p, \mathrm{d}W_V) = 0$. Under this assumption they show that, in a risk-neutral world, the premium C^{HW} on a European call option is the Black and Scholes price C^{BS} evaluated at the average integrated variance $\overline{V}$,

$$\overline{V} = \frac{1}{T-t}\int_t^T V(s)\mathrm{d}s \,, \tag{35}$$

integrated over the distribution $h(\overline{V}|V(t))$ of $\overline{V}$:

$$C^{\mathrm{HW}}(p(t), V(t)) = \int C^{\mathrm{BS}}(\overline{V})h(\overline{V}|V(t))\mathrm{d}\overline{V} \,. \tag{36}$$

The early efforts to identify a more realistic probabilistic model for the underlying return were slowed by the analytical and computational complexity of the option pricing problem. Unlike the BS setting, the early SV specifications do not admit closed-form solutions. Thus, the evaluation of the option price requires time-consuming computations through, e. g., simulation methods or nu-



Stochastic Volatility, Figure 1

Pre- and post-1987 crash implied volatilities. The plots depict Black-Scholes implied volatilities computed from near-maturity options on the S&P 500 futures on October 14, 1987 (the week before the 1987 market crash) and a year later

merical solution of the pricing partial differential equation by finite difference methods. Further, the presence of a latent factor, volatility, and the lack of closed-form expressions for the likelihood function complicate the estimation problem.

Consequently, much effort has gone into developing restrictions for the distribution of the underlying return process that allow for (semi) closed-form solutions and are consistent with the empirical properties of the data. The 'affine' class of continuous-time models has proven particularly useful in providing a flexible, yet analytically tractable, setting. Roughly speaking, the defining feature of affine jump-diffusions is that the drift term, the conditional covariance term, and the jump intensity are all a linear-plus-constant (affine) function of the state vector. The Vasicek [255] bond valuation model and the Cox et al. [112] intertemporal asset pricing model provide powerful examples of the advantages of the affine paradigm.

To illustrate the progress in option pricing applications built on affine models, consider the return dynamics

$$\mathrm{d}p(t) = \mu \mathrm{d}t + \sqrt{V(t)}\mathrm{d}W_p(t) + \xi_p(t)\mathrm{d}q(t) \qquad (37)$$

$$\mathrm{d}V(t) = \kappa(\overline{V} - V(t))\mathrm{d}t + \sigma_V \sqrt{V(t)}\mathrm{d}W_V(t) + \xi_V(t)\mathrm{d}q(t)\,, \qquad (38)$$

where W_p and W_V are standard Brownian motions with non-zero correlation $\rho = \mathrm{corr}(\mathrm{d}W_p, \mathrm{d}W_V)$, q is a Poisson process, uncorrelated with W_p and W_V, with jump intensity

$$\lambda(t) = \lambda_0 + \lambda_1 V(t)\,, \tag{39}$$

that is, $\mathrm{Prob}(\mathrm{d}q_t = 1) = \lambda(t)\mathrm{d}t$. The jump amplitudes variables ξ_p and ξ_V have distributions

$$\xi_V(t) \rightsquigarrow \exp(\overline{\xi}_V) \tag{40}$$

$$\xi_p(t)|\xi_V(t) \rightsquigarrow N\left(\overline{\xi}_p + \rho_\xi \xi_V(t), \sigma_p^2\right)\,. \tag{41}$$

Here volatility is not only stochastic but also subject to jumps which occur simultaneously with jumps in the underlying return process. The Black and Scholes model is a special case of (37)–(41) for constant volatility, $V(t) = \sigma^2, 0 \leq t \leq T$, and no jumps, $\lambda(t) = 0, 0 \leq t \leq T$. The Merton [214] model arises from (37)–(41) if volatility is constant but we allow for jumps in returns.

More recently, Heston [170] has considered a special case of (37)–(41) with stochastic volatility but without jumps. Using transform methods he derives a European option pricing formula which may be evaluated readily through simple numerical integration. His SV model has GARCH-type features, in that the variance is persistent and mean reverts at a rate κ to the long-run mean $\overline{V}$. Compared to Hull and White's [174] setting, Heston's model allows for shocks to returns and volatility to be negatively correlated, i. e., $\rho < 0$, which creates a leverage-type effect and skews the return distribution. This feature is consistent with the properties of equity index returns. Further, a fatter left tail in the return distribution results in a higher cost for crash insurance and therefore makes out-of-the-money put options more expensive. This is qualitatively consistent with the patterns in implied volatilities observed after the 1987 market crash and discussed above.

Bates [56] has subsequently extended Heston's approach to allow for jumps in returns and using similar transform methods he has obtained a semi-closed form solution for the option price. The addition of jumps provides a more realistic description of equity returns and has important option pricing implications. With diffusive shocks (e. g., stochastic volatility) alone a large drop in the value of the underlying asset over a short time span is very unlikely whereas a market crash is always possible as long as large negative jumps can occur. This feature increases the value of a short-dated put option, which offers downside protection to a long position in the underlying asset.

Finally, Duffie et al. [130] have introduced a general model with jumps to volatility which embeds the dynamics (37)–(41). In model (37)–(41), the likelihood of a jump to occur increases when volatility is high ($\lambda_1 > 0$) and a jump in returns is accompanied by an outburst of volatility. This is consistent with what is typically observed during times of market stress. As in the Heston case, variance is persistent with a mean reversion coefficient κ towards its *diffusive* long-run mean $\overline{V}$, while the total long-run variance mean is the sum of the diffusive and jump components. In the special case of constant jump intensity, i. e., $\lambda_1 = 0$, the total long-run mean is $\overline{V} + \overline{\xi}_V \lambda_0/\kappa$. The jump term $(\xi_V(t)\mathrm{d}q(t))$ fattens the right tail of the variance distribution, which induces leptokurtosis in the return distribution. Two effects generate asymmetrically distributed returns. The first channel is the diffusive leverage effect, i. e., $\rho < 0$, the second is the correlation between the volatility and the jump amplitude of returns generated through the coefficient ρ_ξ. Taken together, these effects increase model-implied option prices and help produce a realistic volatility smirk.

Several empirical studies rely on models of the form (37)–(41) in option-pricing applications. For instance, Bates [56] uses Deutsche Mark options to estimate a model with stochastic volatility and constant-intensity jumps to returns, while Bates [57] fits a jump-diffusion model with two SV factors to options on S&P 500 futures. In the latter case, the two SV factors combine to help capture features of the long-run memory in volatility while retaining the analytical tractability of the affine setting (see, e. g., [101] for another model with similar features). Alternative approaches to model long memory in continuous-time SV models rely on the fractional Brownian motion process, e. g., Comte and Renault [108] and Comte et al. [107], while Breidt et al. [76], Harvey [166] and Deo et al. [118] consider discrete-time SV models (see [175] for a review). Bakshi et al. [34,37] estimate a model similar to the one introduced by Bates [56] using S&P 500 options.

Other scholars rely on underlying asset return data alone for estimation. For instance, Andersen et al. [15] and Chernov et al. [95] use equity-index returns to estimate jump-diffusion SV models within and outside the affine (37)–(41) class. Eraker et al. [142] extend this analysis and fit a model that includes constant-intensity jumps to returns and volatility.

Finally, another stream of work examines the empirical implications of SV jump-diffusions using a joint sample of S&P 500 options and index returns. For example, Benzoni [59], Chernov and Ghysels [93], and Jones [190] estimate different flavors of the SV model without jumps. Pan [220] fits a model that has jumps in returns with time-varying intensity, while Eraker [141] extends Pan's work by adding jumps in volatility.

Overall, this literature has established that the SV jump-diffusion model dramatically improves the fit of underlying index returns and options prices compared to the

Black and Scholes model. Stochastic volatility alone has a first-order effect and jumps further enhance model performance by generating fatter tails in the return distribution and reducing the pricing error for short-dated options. The benefits of the SV setting are also significant in hedging applications.

Another aspect related to the specification of SV models concerns the pricing of volatility and jump risks. Stochastic volatility and jumps are sources of uncertainty. It is an empirical issue to determine whether investors demand to be compensated for bearing such risks and, if so, what the magnitude of the risk premium is. To examine this issue it is useful to write model (37)–(41) in so-called risk-neutral form. It is common to assume that the volatility risk premium is proportional to the instantaneous variance, $\eta(t) = \eta_V V(t)$. Further, the adjustment for jump risk is accomplished by assuming that the amplitude $\tilde{\xi}_p(t)$ of jumps to returns has mean $\bar{\tilde{\xi}}_p = \bar{\xi}_p + \eta_p$. These specifications are consistent with an arbitrage-free economy. More general specifications can also be supported in a general equilibrium setting, e. g., a risk adjustment may apply to the jump intensity $\lambda(t)$. However, the coefficients associated to these risk adjustments are difficult to estimate and to facilitate identification they typically are fixed at zero. Incorporating such risk premia in model (37)–(41) yields the following risk-neutral return dynamics (e. g., Pan [220] and Eraker [141]):

$$\mathrm{d}p(t) = (r - \mu^*)\mathrm{d}t + \sqrt{V(t)}\mathrm{d}\widetilde{W}_p(t) + \tilde{\xi}_p(t)\mathrm{d}q(t) \quad (42)$$

$$\begin{aligned}\mathrm{d}V(t) = {}& [\kappa(\overline{V} - V(t)) + \eta_V V(t)]\mathrm{d}t \\ & + \sigma_V \sqrt{V(t)}\mathrm{d}\widetilde{W}_V(t) + \xi_V(t)\mathrm{d}q(t)\,, \end{aligned} \quad (43)$$

where r is the risk-free rate, μ^* a jump compensator term, $\widetilde{W}_p$ and $\widetilde{W}_V$ are standard Brownian motions under this so-called $\mathcal{Q}$ measure, and the risk-adjusted jump amplitude variable $\tilde{\xi}_p$ is assumed to follow the distribution,

$$\tilde{\xi}_p(t)|\xi_V(t) \rightsquigarrow N\left(\bar{\tilde{\xi}}_p + \rho_\xi \xi_V(t), \sigma_p^2\right) . \quad (44)$$

Several studies estimate the risk-adjustment coefficients η_V and η_p for different specifications of model (37)–(44); see, e. g., Benzoni [59], Broadie et al. [78], Chernov and Ghysels [93], Eraker [141], Jones [190], and Pan [220]. It is found that investors demand compensation for volatility and jump risks and these risk premia are important for the pricing of index options. This evidence is reinforced by other studies examining the pricing of volatility risk using less structured but equally compelling procedures. For instance, Coval and Shumway [111] find that the returns on zero-beta index option straddles (i. e., combinations of calls and puts that have offsetting covariances with the index) are significantly lower than the risk-free return. This evidence suggests that in addition to market risk at least a second factor (likely, volatility) is priced in the index option market. Similar conclusions are obtained by Bakshi and Kapadia [36], Buraschi and Jackwerth [79], and Broadie et al. [78].

Risk-Free Bonds and Their Derivatives

The market for (essentially) risk-free Treasury bonds is liquid across a wide maturity spectrum. No-arbitrage restrictions constrain the allowable dynamics in the cross-section of bond yields. Much work has gone into the development of tractable dynamic term structure models capable of capturing the salient time-series properties of interest rates while respecting such cross-sectional no-arbitrage conditions. The class of so-called 'affine' dynamic term structure models provides a flexible and arbitrage-free, yet analytically tractable, setting for capturing the dynamics of the term structure of interest rates. Following Duffie and Kan [129], Dai and Singleton [114,115], and Piazzesi [226], the short term interest rate, $y_0(t)$, is an affine (i. e., linear-plus-constant) function of a vector of state variables, $X(t) = \{x_i(t),\ i = 1, \ldots, N\}$:

$$y_0(t) = \delta_0 + \sum_{i=1}^{N} \delta_i x_i(t) = \delta_0 + \delta_X' X(t)\,, \quad (45)$$

where the state-vector X has risk-neutral dynamics

$$\mathrm{d}X(t) = \tilde{\mathcal{K}}(\tilde{\Theta} - X(t))\mathrm{d}t + \Sigma\sqrt{S(t)}\mathrm{d}\widetilde{W}(t)\,. \quad (46)$$

In Eq. (46), $\widetilde{W}$ is an N-dimensional Brownian motion under the so-called $\mathcal{Q}$-measure, $\tilde{\mathcal{K}}$ and $\tilde{\Theta}$ are $N \times N$ matrices, and $S(t)$ is a diagonal matrix with the ith diagonal element given by $[S(t)]_{ii} = \alpha_i + \beta_i' X(t)$. Within this setting, the time-t price of a zero-coupon bond with time-to-maturity τ is given by

$$P(t, \tau) = \mathrm{e}^{A(\tau) - B(\tau)' X(t)}\,, \quad (47)$$

where the functions $A(\tau)$ and $B(\tau)$ solve a system of ordinary differential equations (ODEs); see, e. g., Duffie and Kan [129]. Semi-closed form solutions are also available for bond derivatives, e. g., bond options as well as caps and floors (e. g., Duffie et al. [130]).

In empirical applications it is important to also establish the evolution of the state vector X under the physical probability measure $\mathcal{P}$, which is linked to the $\mathcal{Q}$-dynamics (46) through a market price of risk, $\Lambda(t)$. Following Dai and Singleton [114] the market price of risk is often given by

$$\Lambda(t) = \sqrt{S(t)}\lambda\,, \tag{48}$$

where λ is an $N \times 1$ vector of constants. More recently, Duffee [127] proposed a broader 'essentially affine' class, which retains the tractability of standard models but, in contrast to the specification in Eq. (48), allows compensation for interest rate risk to vary independently of interest rate volatility. This additional flexibility proves useful in forecasting future yields. Subsequent generalization are in Duarte [124] and Cheridito et al. [92].

Litterman and Scheinkman [201] demonstrate that virtually all variation in US Treasury rates is captured by three factors, interpreted as changes in 'level', 'steepness', and 'curvature'. Consistent with this evidence, much of the term-structure literature has focused on three-factor models. One problem with these models, however, is that the factors are latent variables void of immediate economic interpretation. As such, it is challenging to impose appropriate identifying conditions for the model coefficients and in particular to find the ideal representation for the 'most flexible' model, i. e., the model with the highest number of identifiable coefficients. Dai and Singleton [114] conduct an extensive specification analysis of multi-factor affine term structure models. They classify these models into subfamilies according to the number of (independent linear combination of) state variables that determine the conditional variance matrix of the state vector. Within each subfamily, they proceed to identify the models that lead to well-defined bond prices (a condition they label 'admissibility') and among the admissible specifications they identify a 'maximal' model that nests econometrically all others in the subfamily. Joslin [191] builds on Dai and Singleton's [114] work by pursuing identification through a normalization of the drift term in the state vector dynamics (instead of the diffusion term, as in Dai and Singleton [114]). Duffie and Kan [129] follow an alternative approach to obtain an identifiable model by rotating from a set of latent state variables to a set of observable zero-coupon yields. Collin-Dufresne et al. [104] build on the insights of both Dai and Singleton [114] and Duffie and Kan [129]. They perform a rotation of the state vector into a vector that contains the first few components in the Taylor series expansion of the yield curve around a maturity of zero and their quadratic variation. One advantage is that the elements of the rotated state vector have an intuitive and unique economic interpretation (such as level, slope, and curvature of the yield curve) and therefore the model coefficients in this representation are identifiable. Further, it is easy to construct a model-independent proxy for the rotated state vector, which facilitates model estimation as well as interpretation of the estimated coefficients across models and sample periods.

This discussion underscores an important feature of affine term structure models. The dependence of the conditional factor variance $S(t)$ on one or more of the elements in X introduces stochastic volatility in the yields. However, when a square-root factor is present parametric restrictions (admissibility conditions) need to be imposed so that the conditional variance $S(t)$ is positive over the range of X. These restrictions affect the correlations among the factors which, in turn, tend to worsen the cross-sectional fit of the model. Specifically, CIR models in which $S(t)$ depends on all the elements of X require the conditional correlation among the factors to be zero, while the admissibility conditions imposed on the matrix $\mathcal{K}$ renders the unconditional correlations non-negative. These restrictions are not supported by the data. In contrast, constant-volatility Gaussian models with no square-root factors do not restrict the signs and magnitude of the conditional and unconditional correlations among the factors but they do, of course, not accommodate the pronounced and persistent volatility fluctuations observed in bond yields. The class of models introduced by Dai and Singleton [114] falls between these two extremes. By including both Gaussian and square-root factors they allow for time-varying conditional volatilities of the state variables and yet they do not constrain the signs of some of their correlations. This flexibility helps to address the trade off between generating realistic correlations among the factors while capturing the time-series properties of the yields' volatility.

A related aspect of (unconstrained) affine models concerns the dual role that square-root factors play in driving the time-series properties of yields' volatility and the term structure of yields. Specifically, the time-t yield $y_\tau(t)$ on a zero-coupon bond with time-to-maturity τ is given by

$$P(t,\tau) = \mathrm{e}^{-\tau y_\tau(t)}\,. \tag{49}$$

Thus, we have

$$y_\tau(t) = -\frac{A(\tau)}{\tau} + \frac{B(\tau)'}{\tau}X(t)\,. \tag{50}$$

It is typically assumed that the B matrix has full rank and therefore Eq. (50) provides a direct link between the state-vector $X(t)$ and the term-structure of bond yields. Further, Itô's Lemma implies that the yield y_τ also follows a diffusion process:

$$\mathrm{d}y_\tau(t) = \mu_{y_\tau}(X(t),t)\mathrm{d}t + \frac{B(\tau)'}{\tau}\Sigma\sqrt{S(t)}\mathrm{d}\widetilde{W}(t)\,. \tag{51}$$

Consequently, the (instantaneous) quadratic variation of the yield given as the squared yield volatility coefficient for y_τ is

$$V_{y_\tau}(t) = \frac{B(\tau)'}{\tau} \Sigma S(t) \Sigma' \frac{B(\tau)}{\tau} . \tag{52}$$

The elements of the $S(t)$ matrix are affine in the state vector $X(t)$, i. e., $[S(t)]_{ii} = \alpha_i + \beta_i' X(t)$. Further, invoking the full rank condition on $B(\tau)$, Eq. (50) implies that each state variable in the vector $X(t)$ is an affine function of the bond yields $Y(t) = \{y_{\tau_j}(t), \; j = 1, \ldots, J\}$. Thus, for any τ there is a set of constants $a_{\tau,j}, \; j = 0, \ldots, J$, so that

$$V_{y_\tau}(t) = a_{\tau,0} + \sum_{j=1}^{J} a_{\tau,j} y_{\tau_j}(t) . \tag{53}$$

Hence, the current quadratic yield variation for bonds at any maturity is a linear combination of the term structure of yields. As such, the market is complete, i. e., volatility is perfectly spanned by a portfolio of bonds.

Collin-Dufresne and Goldstein [103] note that this spanning condition is unnecessarily restrictive and propose conditions which ensures that volatility no longer directly enters the main bond pricing Eq. (47). This restriction, which they term 'unspanned stochastic volatility' (USV), effectively breaks the link between the yields' quadratic variation and the level of the term structure by imposing a reduced rank condition on the $B(\tau)$ matrix. Further, since their model is a special (nested) case of the affine class it retains the analytical tractability of the affine model class. Recently Joslin [191] has derived more general conditions for affine term structure models to exhibit USV. His restrictions also produce a market incompleteness (i. e., volatility cannot be hedged using a portfolio of bonds) but do not constrain the degree of mean reversion of the other state variables so that his specification allows for more flexibility in capturing the persistence in interest rate series. (See also the USV conditions in the work by Trolle and Schwartz [253]).

There is conflicting evidence on the volatility spanning condition in fixed income markets. Collin-Dufresne and Goldstein [103] find that swap rates have limited explanatory power for returns on at-the-money 'straddles', i. e., portfolios mainly exposed to volatility risk. Similar findings are in Heidari and Wu [169], who show that the common factors in LIBOR and swap rates explain only a limited part of the variation in the swaption implied volatilities. Moreover, Li and Zhao [197] conclude that some of the most sophisticated multi-factor dynamic term structure models have serious difficulties in hedging caps and cap straddles, even though they capture bond yields well. In contrast, Fan et al. [143] argue that swaptions and even swaption straddles can be well hedged with LIBOR bonds alone, supporting the notion that bond markets are complete.

More recently other studies have examined several versions of the USV restriction, again coming to different conclusions. A direct comparison of these results, however, is complicated by differences in the model specification, the estimation method, and the data and sample period used in the estimation. Collin-Dufresne et al. [105] consider swap rates data and fit the model using a Bayesian Markov Chain Monte Carlo method. They find that a standard three-factor model generates a time series for the variance state variable that is essentially unrelated to GARCH estimates of the quadratic variation of the spot rate process or to implied variances from options, while a four-factor USV model generates both realistic volatility estimates and a good cross-sectional fit. In contrast, Jacobs and Karoui [178] consider a longer data set of US Treasury yields and pursue quasi-maximum likelihood estimation. They find the correlation between model-implied and GARCH volatility estimates to be high. However, when estimating the model with a shorter sample of swap rates, they find such correlations to be small or negative. Thompson [250] explicitly tests the Collin-Dufresne and Goldstein [103] USV restriction and rejects it using swap rates data. Bikbov and Chernov [62], Han [164], Jarrow et al. [183], Joslin [192], and Trolle and Schwartz [254] rely on data sets of derivatives prices and underlying interest rates to better identify the volatility dynamics.

Andersen and Benzoni [12] directly relate model-free realized volatility measures (constructed from high-frequency US Treasury data) to the cross-section of contemporaneous bond yields. They find that the explanatory power of such regressions is very limited, which indicates that volatility is not spanned by a portfolio of bonds. The evidence in Andersen and Benzoni [12] is consistent with the USV models of Collin-Dufresne et al. [105] and Joslin [191], as well as with a model that embeds weak dependence between the yields and volatility as in Joslin [192]. Moreover, Duarte [125] argues that the effects of mortgage-backed security hedging activity affects both the interest rate volatility implied by options and the actual interest rate volatility. This evidence suggests that variables that are not in the span of the term structure of yields and forward rates contribute to explain volatility in fixed income markets. Also related, Wright and Zhou [258] find that adding a measure of market jump volatility risk to a regression of excess bond returns on the term structure of forward rates nearly doubles the R^2 of the regression.

Taken together, these findings suggest more generally that genuine SV models are critical for appropriately capturing the dynamic evolution of the term structure.

Estimation Methods

There are a very large number of alternative approaches to estimation and inference for parametric SV models and we abstain from a thorough review. Instead, we point to the basic challenges that exist for different types of specifications, how some of these were addressed in the early literature and finally provide examples of methods that have been used extensively in recent years. Our exposition continues to focus on applications to equity returns, interest rates, and associated derivatives.

Many of the original SV models were cast in discrete time, inspired by the popular GARCH paradigm. In that case, the distinct challenge for SV models is the presence of a strongly persistent latent state variable. However, more theoretically oriented models, focusing on derivatives applications, were often formulated in continuous time. Hence, it is natural that the econometrically-oriented literature has moved in this direction in recent years as well. This development provides an added complication as the continuous-time parameters must be estimated from discrete return data and without direct observations on volatility. For illustration, consider a fully parametric continuous-time SV model for the asset return r with conditional variance V and coefficient vector Ψ. Most methods to estimate Ψ rely on the conditional density f for the data generating process,

$$f(r(t), V(t)|\mathcal{I}(t-1), \Psi) = f_{r|V}(r(t)|V(t), \mathcal{I}(t-1), \Psi) \times f_V(V(t)|\mathcal{I}(t-1), \Psi) , \quad (54)$$

where $\mathcal{I}(t-1)$ is the available information set at time $t-1$. The main complications are readily identified. First, analytic expressions for the discrete-time transition (conditional) density, f, or the discrete-time moments implied by the data generating process operating in continuous time, are often unavailable. Second, volatility is latent in SV models, so that even if a closed-form expression for f is known, direct evaluation of the above expression is infeasible due to the absence of explicit volatility measures. The marginal likelihood with respect to the observable return process alone is obtained by integrating over all possible paths for the volatility process, but this integral has a dimension corresponding to sample size, rendering the approach infeasible in general.

Similar issues are present when estimating continuous-time dynamic term structure models. Following Piazzesi [227], a change of variable gives the conditional density for a zero-coupon yield y on a bond with time to maturity τ:

$$f(y_\tau(t)|\mathcal{I}(t-1), \Psi) = f_X(g(y_\tau(t), \Psi)|\mathcal{I}(t-1), \Psi) \times |\nabla_y g(y_\tau(t), \Psi)| . \quad (55)$$

Here the latent state vector X has conditional density f_X, the function $g(\cdot, \Psi)$ maps the observable yield y into X, $X(t) = g(y_\tau(t), \Psi)$, and $\nabla_y g(y_\tau(t), \Psi)$ is the Jacobian determinant of the transformation. Unfortunately, analytic expressions for the conditional density f_X are known only in some special cases. Further, the mapping $X(t) = g(y_\tau(t), \Psi)$ holds only if the model provides an exact fit to the yields, while in practice different sources of error (e. g., model mis-specification, microstructure effects, measurement errors) inject a considerable degree of noise into this otherwise deterministic linkage (for correct model specification) between the state vector and the yields. As such, a good measure of X might not be available to evaluate the conditional density (55).

Estimation via Discrete-Time Model Specification or Approximation

The first empirical studies have estimated discrete-time SV models via a (Generalized) Method of Moments procedure by matching a number of theoretical and sample moments, e. g., Chan et al. [89], Ho et al. [171], Longstaff and Schwartz [204], and Melino and Turnbull [211]. These models were either explicitly cast in discrete time or were seen as approximate versions of the continuous-time process of interest. Similarly, several authors estimate diffusive affine dynamic term structure models by approximating the continuous-time dynamics with a discrete-time process. If the error terms are stipulated to be normally distributed, the transition density of the discretized process is multivariate normal and computation of unconditional moments then only requires knowledge of the first two moments of the state vector. This result facilitates quasi-maximum likelihood estimation. In evaluating the likelihood function, some studies suggest using closed-form expressions for the first two moments of the continuous-time process instead of the moments of the discretized process (e. g., Fisher and Gilles [145] and Duffee [127]), thus avoiding the associated discretization bias. This approach typically requires some knowledge of the state of the system which may be obtained, imperfectly, through matching the system, given the estimated parameter vector, to a set of observed zero-coupon yields to infer the state vector X. A modern alternative is to use the so-called particle filter as an efficient filtering procedure for the unobserved state variables given the estimated parameter vector. We provide more detailed accounts of both of these procedures later in this section.

Finally, a number of authors develop a simulated maximum likelihood method that exploit the specific structure of the discrete-time SV model. Early examples are Danielsson and Richard [117] and Danielsson [116] who exploit the Accelerated Gaussian Importance Sampler for efficient Monte Carlo evaluation of the likelihood. Subsequent improvements were provided by Fridman and Harris [149] and Liesenfeld and Richard [200], with the latter relying on Efficient Importance Sampling (EIS). In a second step, EIS can also be used for filtering the latent volatility state vector. In general, these inference techniques provide quite impressive efficiency but the methodology is not always easy to generalize beyond the structure of the basic discrete-time SV asset return model. We discuss the general inference problem for continuous-time SV models for which the lack of a closed-form expression for the transition density is an additional complicating factor in a later section.

Filtering the Latent State Variable Directly During Estimation

Some early studies focused on direct ways to extract estimates of the latent volatility state variable in discrete-time SV asset return models. The initial approach was based on quasi-maximum likelihood (QML) methods exploiting the Kalman filter. This method requires a transformation of the SV model to a linear state-space form. For instance, Harvey and Shephard [168] consider a version of the Taylor's [249] discrete-time SV model,

$$p(t) = p(t-1) + \beta + \sqrt{V(t)}\varepsilon(t) \tag{56}$$

$$\log(V(t)) = \alpha + \phi \log(V(t-1)) + \eta(t)\,, \tag{57}$$

where p is the logarithmic price, ε is a zero-mean error term with unit variance, and η is an independently-distributed error term with zero mean and variance σ_η^2.

Define $y(t) = p(t) - p(t-1) - \beta$, square the observations in Eq. (56), and take logarithms to obtain the *measurement equation*,

$$\ell(t) = \omega + h(t) + \xi(t)\,, \tag{58}$$

where $\ell(t) \equiv \log y(t)^2$, $h(t) \equiv \log(V(t))$. Further, ξ is a zero-mean disturbance term given by $\xi(t) = \log(\varepsilon(t)^2) - \mathrm{E}[\log(\varepsilon(t)^2)]$, $\omega = \log(\sigma^2) + \mathrm{E}[\log(\varepsilon(t)^2)]$, and σ is a scale constant which subsumes the effect of the drift term α in Eq. (57). The autoregression (57) yields the *transition equation*,

$$h(t) = \phi h(t-1) + \eta(t)\,, \tag{59}$$

Taken together, Eqs. (58) and (59) are the linear state-space transformation of the SV model (56)–(57). If the joint distribution of ε and η is symmetric, i. e., $f(\varepsilon, \eta) = f(-\varepsilon, -\eta)$, then the disturbance terms in the state-space form are uncorrelated even if η and ε are not. A possible dependence between ε and η allows the model to pick up some of the asymmetric behavior often observed in stock returns. Projection of $[h(t) - \mathrm{E}_{t-1} h(t)]$ over $[\,\ell(t) - \mathrm{E}_{t-1}\,\ell(t)\,]$ yields the Kalman filter estimate of the latent (logarithmic) variance process:

$$\begin{aligned} \mathrm{E}_t\, h(t) = \mathrm{E}_{t-1}\, h(t) & \\ &+ \frac{\mathrm{E}\{[h(t) - \mathrm{E}_{t-1}\, h(t)] \times [\ell(t) - \mathrm{E}_{t-1}\,\ell(t)]\}}{\mathrm{E}\{[\ell(t) - \mathrm{E}_{t-1}\,\ell(t)]^2\}} \\ &\times [\ell(t) - \mathrm{E}_{t-1}\,\ell(t)]\,, \end{aligned} \tag{60}$$

where the conditional expectations $\mathrm{E}_{t-1}\,\ell(t)$ and $\mathrm{E}_{t-1}\, h(t)$ are given by:

$$\mathrm{E}_{t-1}\,\ell(t) = \omega + \mathrm{E}_{t-1}\, h(t) \tag{61}$$

$$\mathrm{E}_{t-1}\, h(t) = \phi\, \mathrm{E}_{t-1}\, h(t-1)\,. \tag{62}$$

To start the recursion (60)–(62), the initial value $\mathrm{E}_0\, h(0)$ is fixed at the long-run mean $\log(\overline{V})$.

Harvey and Shephard [168] estimate the model coefficients via quasi-maximum likelihood, i. e. by treating the errors ξ and η as though they were normal and maximizing the prediction-error decomposition form of the likelihood function obtained via the Kalman filter. Inference is valid as long as the standard errors are appropriately adjusted. In their application they rely on daily returns on the value-weighted US market index over 1967–1987 and daily returns for 30 individual stocks over 1974–1983. Harvey et al. [167] pursue a similar approach to fit a multivariate SV model to a sample of four exchange rate series from 1981 to 1985. One major drawback of the Kalman filter approach is that the finite sample properties can be quite poor because the error term, ξ, is highly non-Gaussian, see, e. g., Andersen, Chung, and Sørensen [27]. The method may be extended to accommodate various generalizations including long memory persistence in volatility as detailed in Ghysels, Harvey, and Renault [158].

A related literature, often exploited in multivariate settings, specifies latent GARCH-style dynamics for a state vector which governs the systematic evolution of a higher dimensional set of asset returns. An early representative of these specifications is in Diebold and Nerlove [120], who exploit the Kalman filter for estimation, while Fiorentina et al. [144] provide a likelihood-based estimation procedure using MCMC techniques. We later review the MCMC approach and the associated filtering application, e. g, the 'particle filter', in some detail.

The state-space form is also useful to characterize the dynamics of interest rates. Following, e. g., Piazzesi [226], for a discrete-time dynamic term structure model the measurement and transition equations are

$$y_\tau(t) = -\frac{A(\tau)}{\tau} + \frac{B(\tau)'}{\tau} X(t) + \xi_\tau(t) \tag{63}$$

$$X(t) = \mu + \Phi X(t-1) + \Sigma \sqrt{S(t)}\, \varepsilon(t)\,, \tag{64}$$

where $S(t)$ is a matrix whose elements are affine functions of the state vector X, and A and B solve a system of difference equations. When all the yields are observed with error (i. e., $\xi_\tau \neq 0 \forall \tau,\ 0 \leq \tau \leq T$), QML estimation of the system (63)–(64) via the extended Kalman filter method yields an estimate of the coefficient vector. Applications of this approach for the US term structure data include Campbell and Viceira [81], Gong and Remolona [161], and Pennacchi [225]. The extended Kalman filter involves a linear approximation of the relation between the observed data and the state variables, and the associated approximation error will produce biased estimates. Christoffersen et al. [99] raise this concern and recommend the use of the so-called unscented Kalman filter for estimation of systems in which the relation between data and state variables is highly non-linear, e. g., options data.

Methods Accommodating the Lack of a Closed-Form Transition Density

We have so far mostly discussed estimation techniques for models with either a known transition density or one that is approximated by a discrete-time system. However, the majority of empirically-relevant continuous-time models do not possess explicit transition densities and alternative approaches are necessary. This problem leads us naturally towards the large statistics and econometric literature on estimation of diffusions from discretely-observed data. The vast majority of these studies assume that all relevant variables are observed so the latent volatility or yield curve state variables, integral to SV models, are not accounted for. Nonetheless, it may be feasible to extract the requisite estimates of the state variable by alternate means, thus restoring the feasibility, albeit not efficiency, of the basic approach. Since the literature is large and not directly geared towards genuine SV models, we focus on methods that have seen use in applications involving latent state variables.

A popular approach is to invert the map between the state vector and a subset of the observables assuming that the model prices specific securities exactly. In applications to equity markets this is done, e. g., by assuming that one option contract is priced without error, which implies a specific value (estimate) of the variance process given the model parameters Ψ. For instance, Pan [220] follows this approach in her study of S&P 500 options and returns, which we review in more detail in Sect. "Estimation from Option Data". In applications to fixed income markets it is likewise stipulated that certain bonds are priced without error, i. e., in Eq. (63) the error term $\xi_{\tau_i}(t)$ is fixed at zero for a set of maturities $\tau_1, \ldots, \tau_N$, where N matches the dimension of the state vector X. This approach yields an estimate for the latent variables through the inverse-map $X(t) = g(y_\tau(t), \Psi)$.

One criticism of the state vector inversion procedure is that it requires ad hoc assumptions regarding the choice of the securities that are error-free (those used to compute model-implied measures of the state vector) vis-a-vis those observed with error (used either for estimation or to assess model performance in an 'out-of-sample' cross-sectional check). In fact, the extracted state vector can be quite sensitive to the choice of derivatives (or yields) used. Nevertheless, this approach has intuitive appeal. Model-implied measures of the state vector, in combination with a closed-form expression for the conditional density (55), allow for efficient estimation of the coefficient vector Ψ via maximum likelihood. Analytic expressions for f_X in Eq. (55) exist in a limited number of cases. For instance, if X is Gaussian then f_X is multivariate normal, while if X follows a square-root process then f_X can be expressed in terms of the modified Bessel function (e. g., [113]). Different flavors of these continuous-time models are estimated in, e. g., [91,106,132,182,223]. In more general cases, including affine processes that combine Gaussian and square-root state variables, closed-form expressions for f_X are no longer available. In the rest of this section we briefly review different methods to overcome this problem. The interested reader may consult, e. g., [226] for more details.

Lo [203] warns that the common approach of estimating parameters of an Itô process by applying maximum likelihood to a discretization of the stochastic differential equation yields inconsistent estimators. In contrast, he characterizes the likelihood function as a solution to a partial differential equation. The method is very general, e. g., it applies not only to continuous-time diffusions but also to jump processes. In practice, however, analytic solutions to the partial differential equations (via, e. g., Fourier transforms) are available only for a small class of models so computationally-intensive methods (e. g., finite differencing or simulations) are generally required to solve the problem. This is a severe limitation in the case of multivariate systems like SV models.

For general Markov processes, where the above solution is infeasible, a variety of procedures have been advocated in recent years. Three excellent surveys provide different perspectives on the issue. Aït-Sahalia, Hansen, and Scheinkman [5] discuss operator methods and mention the potential of applying a time deformation technique to account for genuine SV features of the process, as in Conley, Hansen, Luttmer, and Scheinkman [109]. In addition, the Aït-Sahalia [3,4] closed-form polynomial expansions for discretely-sampled diffusions are reviewed along with the Schaumburg [235] extension to a general class of Markov processes with Lévy-type generators. Meanwhile, Bibby, Jacobsen, and Sørensen [61] survey the extensive statistics literature on estimating functions for diffusion-type models and Bandi and Phillips [42] explicitly consider dealing with nonstationary processes (see also the work of Bandi [39], Bandi and Nguyen [41], and Bandi and Phillips [43,44]).

The characteristic function based inference technique has been particularly widely adopted due to the natural fit with the exponentially affine model class which provides essentially closed-form solutions for many pricing applications. Consequently, we dedicate a separate section to this approach.

Characteristic Functions Singleton [242] proposes to exploit the information contained in the conditional characteristic function of the state vector X,

$$\phi(iu, X(t), \Psi) = \mathrm{E}\left[\mathrm{e}^{iu' X(t+1)} \middle| X(t)\right] , \qquad (65)$$

to pursue maximum likelihood estimation of affine term structure models. In Equation (65) we highlight the dependence of the characteristic function on the unknown parameter vector Ψ. When X is an affine (jump-)diffusion process, ϕ has the exponential affine form,

$$\phi(iu, X(t), \Psi) = \mathrm{e}^{\alpha_t(u) + \beta_t(u)' X(t)} , \qquad (66)$$

where the functions α and β solve a system of ODEs. As such, the transition density f_X is known explicitly up to an inverse-Fourier transformation of the characteristic function (65),

$$f_X(X(t+1) \big| X(t); \Psi) = \frac{1}{\pi^N} \int_{\mathbb{R}^N_+} \mathrm{Re}\left[\mathrm{e}^{-iu' X(t+1)} \phi(iu, X(t), \Psi)\right] \mathrm{d}u \, . \qquad (67)$$

Singleton shows that Gauss–Legendre quadrature with a relatively small number of quadrature points allows to accurately evaluate the integral in Eq. (67) when X is univariate. As such, the method readily delivers efficient estimates of the parameter vector, Ψ, subject to an auxiliary assumption, namely that the state vector may be extracted by assuming that a pre-specified set of security prices is observed without error while the remainder have non-trivial error terms.

When X is multivariate the Fourier inversion in Eq. (67) is computationally more demanding. Thus, when estimating multi-dimensional systems Singleton suggests focusing on the conditional density function of the individual elements of X, but conditioned on the full state vector,

$$f_{X_j}(X_j(t+1) | X(t); \Psi) = \frac{1}{2\pi} \int_{\mathbb{R}} \mathrm{e}^{-i\omega \mathbf{I}_j' X(t+1)} \phi(i\omega \mathbf{I}_j, X(t), \Psi) \mathrm{d}\omega \, , \qquad (68)$$

where the vector $\mathbf{I}_j$ has 1 in the jth element and zero elsewhere so that the jth element of X is $X_j(t+1) = \mathbf{I}_j' X(t+1)$. Maximization of the likelihood function obtained from f_{X_j}, for a fixed j, will often suffice to obtain a consistent estimate of Ψ. Exploiting more than one of the conditional densities (68) will result in more efficient Ψ estimate. For instance, the scores of multiple univariate log-likelihood functions, stacked in a vector, yield moment conditions that allow for generalized method of moment (GMM) estimation of the system. Alternatively, Joslin [192] proposes a change-of-measure transformation which reduces the oscillatory behavior of the integrand in Eq. (67). When using this transformation, Gauss-Hermite quadrature more readily provides a solution to the integral in (67) even if the state vector X is multi-dimensional, thus facilitating full ML estimation of the system.

Related, several studies have pursued GMM estimation of affine processes using characteristic functions. Definition (65) yields the moment condition

$$\mathrm{E}\left[(\phi(iu, X(t), \Psi) - \mathrm{e}^{iu' X(t+1)}) z(u, X(t))\right] = 0 \, , \qquad (69)$$

where X is an N-dimensional (jump-)diffusion, $u \in \mathbb{R}^N$, and z is an instrument function. When X is affine, the characteristic function takes the exponential form (66). Different choices of u and z yield a set of moment conditions that can be used for GMM estimation and inference. Singleton [242] derives the optimal instrument in terms of delivering efficient estimates. Carrasco et al. [86] approximate the optimal instrument with a set of basis functions that do not require the knowledge of the conditional likelihood function, thus avoiding one of the assumptions invoked by Singleton. Further, they build on Carrasco and Florens [87] to implement estimation using a continuum of moment conditions, which yields maximum-likelihood efficiency. Other applications of GMM-characteristic function methods to affine (jump-) diffusions for equity index returns are in Chacko and Viceira [88] and Jiang and Knight [184].

In some cases the lack of closed-form expressions for the moment condition in Eq. (69) can hinder GMM estimation. In these cases the expectation in Eq. (69) can be evaluated by Monte Carlo integration. This is accomplished by simulating a long sample from the discretized process for a given value of the coefficient vector Ψ. The parameter Ψ is then estimated via the simulated method of moments (SMM) of McFadden [206] and Duffie and Singleton [131]. Singleton [242] proposes SMM characteristic function estimators that exploit the special structure of affine term structure models.

Efficient Estimation of General Continuous-Time Processes

A number of recent approaches offer excellent flexibility in terms of avoiding approximations to the continuous-time model-implied transition density while still facilitating efficient estimation of the evolution of the latent state vector for the system.

Maximum Likelihood with Characteristic Functions
Bates [58] develops a filtration-based maximum likelihood estimation method for affine processes. His approach relies on Bayes' rule to recursively update the joint characteristic function of latent variables and data conditional on past data. He then obtains the transition density by Fourier inversion of the updated characteristic function.

Denote with $y(t)$ and $X(t)$ the time-t values of the observable variable and the state vector, respectively, and let $Y(t) \equiv \{y(1), \dots, y(t)\}$ be the data observed up to time t. Consider the case in which the characteristic function of $z(t+1) \equiv (y(t+1), X(t+1))$ conditional on $z(t) \equiv (y(t), X(t))$, is an exponential affine function of $X(t)$:

$$\begin{aligned}\phi(is, iu, z(t), \Psi) &= \mathrm{E}\left[\mathrm{e}^{is' y(t+1)+iu' X(t+1)} \middle| z(t)\right] \\ &= \mathrm{e}^{\alpha(is,iu,y(t))+\beta(is,iu,y(t))' X(t)} . \quad (70)\end{aligned}$$

Next, determine the value of the characteristic function conditional on the observed data $Y(t)$:

$$\begin{aligned}&\phi(is, iu, Y(t), \Psi) \\ &\quad = \mathrm{E}\left[\mathrm{E}\left[\mathrm{e}^{is' y(t+1)+iu' X(t+1)} \middle| z(t)\right] \middle| Y(t)\right] \\ &\quad = \mathrm{E}\left[\mathrm{e}^{\alpha(is,iu,y(t))+\beta(is,iu,y(t))' X(t)} \middle| Y(t)\right] \\ &\quad = \mathrm{e}^{\alpha(is,iu,y(t))} \psi(\beta(is, iu, y(t)), Y(t), \Psi) , \quad (71)\end{aligned}$$

where $\psi(iu, Y(t), \Psi) \equiv \mathrm{E}\left[\mathrm{e}^{iu' X(t)} \middle| Y(t)\right]$ denotes the (marginal) characteristic function for the state vector conditional on the observed data. Fourier inversion then yields the conditional density for the observation $y(t+1)$ conditional on $Y(t)$:

$$\begin{aligned}&f_y(y(t+1)|Y(t); \Psi) \\ &\quad = \frac{1}{2\pi} \int_{\mathbb{R}} \mathrm{e}^{-is' y(t+1)} \phi(is, 0, Y(t), \Psi) \mathrm{d}s . \quad (72)\end{aligned}$$

The next step updates the characteristic function ψ (Bartlett [48]):

$$\begin{aligned}\psi(iu, Y(t+1), \Psi) &= \frac{1}{2\pi f_y(y(t+1)|Y(t); \Psi)} \\ &\quad \cdot \int_{\mathbb{R}} \mathrm{e}^{-is' y(t+1)} \phi(is, iu, Y(t), \Psi) \mathrm{d}s . \quad (73)\end{aligned}$$

To start the recursion, Bates initializes ψ at the unconditional characteristic function of the latent variable X. The log-likelihood function is then given by

$$\begin{aligned}\log \mathcal{L}(Y(T); \Psi) &= \log(f_y(y(1); \Psi)) \\ &\quad + \sum_{t=2}^{T} \log(f_y(y(t)|Y(t-1); \Psi)) . \quad (74)\end{aligned}$$

A nice feature is that the method provides a natural solution to the filtering problem. The filtered estimate of the latent state X and its variance are computed from the first and second derivatives of the moment generating function $\psi(u, Y(t); \Psi)$ in Eq. (73), evaluated at $u = 0$:

$$\begin{aligned}\mathrm{E}[X(t+1)|Y(t+1); \Psi] &= \frac{1}{2\pi f_y(y(t+1)|Y(t); \Psi)} \\ &\quad \times \int_{\mathbb{R}} \mathrm{e}^{-is' y(t+1)} \phi_u(is, 0, Y(t); \Psi) \mathrm{d}s \quad (75)\end{aligned}$$

$$\operatorname{Var}[X(t+1)|Y(t+1);\Psi] = \frac{1}{2\pi f_y(y(t+1)|Y(t);\Psi)} \times \int_{\mathbb{R}} \mathrm{e}^{-is'y(t+1)} \phi_{uu}(is, 0, Y(t);\Psi)\mathrm{d}s - \{\mathrm{E}[X(t+1)|Y(t+1)]\}^2 \,. \quad (76)$$

A drawback is that at each step t of the iteration the method requires storage of the entire characteristic function $\psi(iu, Y(t);\Psi)$. To deal with this issue Bates recommends to approximate the true ψ with the characteristic function of a variable with a two-parameter distribution. The choice of the distribution depends on the X-dynamics while the two parameters of the distribution are determined by the conditional mean $\mathrm{E}[X(t+1)|Y(t+1);\Psi]$ and variance $\operatorname{Var}[X(t+1)|Y(t+1);\Psi]$ given in Equations (75)–(76).

In his application Bates finds that the method is successful in estimating different flavors of the SV jump-diffusion for a univariate series of daily 1953–1996 S&P 500 returns. In particular, he shows that the method obtains estimates that are equally, if not more, efficient compared to the efficient method of moments and Markov Chain Monte Carlo methods described below. Extensions of the method to multivariate processes are theoretically possible, but they require numerical integration of multi-dimensional functions, which is computationally demanding.

Simulated Maximum Likelihood In Sect. "Filtering the Latent State Variable Directly During Estimation" we discussed methods for simulated ML estimation and inference in discrete-time SV models. Pedersen [224] and Santa-Clara [234] independently develop a simulated maximum likelihood (SML) method to estimate continuous-time diffusion models. They divide each interval in between two consecutive data points X_{t+1} and X_t into M sub-intervals of length $\Delta = 1/M$ and they discretize the X process using the Euler scheme,

$$X_{t+(i+1)\Delta} = X_{t+i\Delta} + \mu(X_{t+i\Delta})\Delta + \Sigma(X_{t+i\Delta})\sqrt{\Delta}\varepsilon_{t+(i+1)\Delta} \,, \quad i = 0, \ldots, M-1 \,, \quad (77)$$

where μ and Σ are the drift and diffusion terms of the X process and ε is multivariate normal with mean zero and identity variance matrix. The transition density of the discretized process is multivariate normal with mean μ and variance matrix $\Sigma\Sigma'$. As Δ goes to zero, this density converges to that of the continuous-time process X. As such, the transition density from X_t to X_{t+1} is given by

$$f_X(X_{t+1}|X_t;\Psi) = \int f_X(X_{t+1}|X_{t+1-\Delta};\Psi) \times f_X(X_{t+1-\Delta}|X_t;\Psi)\mathrm{d}X_{t+1-\Delta} \,. \quad (78)$$

For sufficiently small values of Δ the first term in the integrand, $f_X(X_{t+1}|X_{t+1-\Delta};\Psi)$, is approximated by the transition density of the discretized process, while the second term, $f_X(X_{t+1-\Delta}|X_t;\Psi)$, is a multi-step-ahead transition density that can be computed from the recursion from X_t to $X_{t+1-\Delta}$. Writing the right-hand side of Eq. (78) as a conditional expectation yields

$$f_X(X_{t+1}|X_t;\Psi) = E_{X_{t+1-\Delta}|X_t}\left[f_X(X_{t+1}|X_{t+1-\Delta};\Psi)\right] . \quad (79)$$

The expectation in Eq. (79) can be computed by Monte Carlo integration over a large number of paths for the process X, simulated via the Euler scheme (77). As Δ vanishes, the Euler scheme is consistent. Thus, when the size of the simulated sample increases the sample average of the function f_X, evaluated at the random draws of $X_{t+1-\Delta}$, converges to the true transition density. Application of the principles in Bladt and Sørensen [64] may well be useful in enhancing the efficiency of the simulation scheme and hence the actual efficiency of the inference procedure in practice.

Brandt and Santa-Clara [75] apply the SML method to estimate a continuous-time model of the joint dynamics of interest rates in two countries and the exchange rate between the two currencies. Piazzesi [227] extends the SML approach for jump-diffusion processes with time-varying jump intensity. She considers a high-frequency policy rule based on yield curve information and an arbitrage-free bond market and estimates the model using 1994–1998 data on the Federal Reserve target rate, the six-month LIBOR rate, and swap yields.

An important issue is how to initialize any unobserved component of the state vector, $X(t)$, such as the volatility state at each observation to provide a starting point for the next Monte Carlo integration step. This may be remedied through application of the particle filter, as mentioned earlier and discussed below in connection with MCMC estimation. Another possibility is, as also indicated previously, to extract the state variable through inversion from derivatives prices or yields assumed observed without pricing errors.

Indirect Inference There are also other method-of-moments strategies to estimate finitely-sampled continuous-time processes of a general type. One prominent approach approximates the unknown transition density for the continuous-time process with the density of a semi-nonparametric (SNP) auxiliary model. Then one can use the score function of the auxiliary model to form moment conditions for the parameter vector $\varPsi$ of the continuous-time model. This approach yields the efficient method of moments estimator (EMM) of Gallant and Tauchen [154], Gallant et al. [150], and Gallant and Long [152], and the indirect inference estimator of Gouriéroux et al. [162] and Smith [243].

To fix ideas, suppose that the conditional density for a continuous-time return process r (the 'structural' model) is unknown. We intend to approximate the unknown density with a discrete-time model (the 'auxiliary' model) that is tractable and yet sufficiently flexible to accommodate the systematic features of the actual data sample well. A parsimonious auxiliary density for r embeds ARMA and EGARCH leading terms to capture the conditional mean and variance dynamics. There may be residual excess skewness and kurtosis that elude the ARMA and EGARCH forms. As such, the auxiliary density is rescaled using a nonparametric polynomial expansion of order K, which yields

$$g_K(r(t)|x(t);\xi) = \left(\nu + (1-\nu) \times \frac{[P_K(z(t),x(t))]^2}{\int_{\mathbb{R}}[P_K(z(t),x(t))]^2\phi(u)\mathrm{d}u}\right)\frac{\phi(z(t))}{\sqrt{h(t)}}\,, \tag{80}$$

where ν is a small constant, $\phi(.)$ is the standard normal density, $x(t)$ contains lagged return observations, and

$$z(t) = \frac{r(t)-\mu(t)}{\sqrt{h(t)}}\,, \tag{81}$$

$$\mu(t) = \phi_0 + ch(t) + \sum_{i=1}^{s}\phi_i r(t-1) + \sum_{i=1}^{u}\delta_i\varepsilon(t-1)\,, \tag{82}$$

$$\begin{aligned}\log h(t) &= \omega + \sum_{i=1}^{p}\beta_i \log h(t-1) \\ &\quad + (1+\alpha_1 L + \cdots + \alpha_q L^q) \\ &\quad \times \left[\theta_1 z(t-1) + \theta_2(b(z(t-1)) - \sqrt{2/\pi})\right],\end{aligned} \tag{83}$$

$$P_K(z,x) = \sum_{i=0}^{K_z} a_i(x) z^i = \sum_{i=0}^{K_z}\left(\sum_{|j|=0}^{K_x} a_{ij}x^j\right) z^i\,, \quad a_{00} = 1\,. \tag{84}$$

Here j is a multi-index vector, $x^j \equiv (x_1^{j_1},\dots,x_M^{j_M})$, and $|j| \equiv \sum_{m=1}^{M} j_m$. The term $b(z)$ is a smooth (twice-differentiable) function that closely approximates the absolute value operator in the EGARCH variance equation.

In practice, the representation of P_K is given by Hermite orthogonal polynomials. When the order K of the expansion increases, the auxiliary density will approximate the data arbitrarily well. If the structural model is indeed the true data generating process, then the auxiliary density will converge to that of the structural model. For a given K, the QML estimator $\hat{\xi}$ for the auxiliary model coefficient satisfies the score condition

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\partial \log g_K(r(t)|x(t);\hat{\xi})}{\partial\xi} = 0\,. \tag{85}$$

Suppose now that the structural model is correct and $\varPsi_0$ is the true value of its coefficient vector. Consider a series $\{r(t;\varPsi), x(t;\varPsi)\}$, $t = 1,\dots,\mathcal{T}(T)$, simulated from the structural model. Then we expect that the score condition (85) holds when evaluated by averaging over the simulated returns rather than over the actual data:

$$m_{\mathcal{T}(T)}(\varPsi_0,\hat{\xi}) = \frac{1}{\mathcal{T}(T)}\sum_{t=1}^{\mathcal{T}(T)}\frac{\partial \log g_K(r(t,\varPsi_0)|x(t,\varPsi_0);\hat{\xi})}{\partial\xi} \approx 0\,. \tag{86}$$

When T and $\mathcal{T}(T)$ tend to infinity, condition (86) holds exactly.

Gallant and Tauchen [154] propose the EMM estimator $\hat{\varPsi}$ defined via

$$\hat{\varPsi} = \arg\min_{\varPsi} m_{\mathcal{T}(T)}(\varPsi,\hat{\xi})'\,\hat{W}_T m_{\mathcal{T}(T)}(\varPsi,\hat{\xi})\,, \tag{87}$$

where the weighting matrix $\hat{W}_T$ is a consistent estimate of the inverse asymptotic covariance matrix of the auxiliary score function, e. g., the inverse outer product of the SNP gradient:

$$\hat{W}_T^{-1} = \frac{1}{T}\sum_{t=1}^{T}\left[\frac{\partial \log g_K(r(t)|x(t);\hat{\xi})}{\partial\xi}\right] \times \left[\frac{\partial \log g_K(r(t)|x(t);\hat{\xi})}{\partial\xi}\right]'\,. \tag{88}$$

An important advantages of the technique is that EMM estimates achieve the same degree of efficiency as the ML procedure, when the score of the auxiliary model asymptotically spans the score of the true model. It also delivers powerful specification diagnostics that provide guidance in the model selection. Gallant and Tauchen [154] show that the EMM estimator is asymptotically normal. Further, under the assumption that the structural model is correctly specified, they derive a χ^2 statistic for the test of over-identifying restrictions. Gallant et al. [150] normalize the vector $m_{\mathcal{T}(T)}(\hat{\Psi}, \hat{\xi})$ by its standard error to obtain a vector of score t-ratios. The significance of the individual score elements is often informative of the source of model mis-specification, with the usual caveat that failure to capture one characteristic of the data may result in the significance of a moment condition that pertains to a coefficient not directly related to that characteristic (due to correlation in the moment conditions). Finally, EMM provides a straightforward solution to the problem of filtering and forecasting the latent return variance process V, i. e., determining the conditional densities $f(V(t)|x(t), \Psi)$ and $f(V(t+j)|x(t), \Psi)$, $j \geq 0$. This is accomplished through the *reprojection* method discussed in, e. g., Gallant and Long [152] and Gallant and Tauchen [155]. In applications to dynamic term structure models, the same method yields filtered and forecasted values for the latent state variables.

The reprojection method assumes that the coefficient vector Ψ is known. In practice, Ψ is fixed at the EMM estimate $\hat{\Psi}$. Then one simulates a sample of returns and latent variables from the structural model and fits the auxiliary model on the simulated data. This is equivalent to the first step of the EMM procedure except that, in the reprojection step, we fit the auxiliary model assuming the structural model is correct, rather than using actual data. The conditional density of the auxiliary model, estimated under the null, approximates the unknown density of the structural model:

$$g_K(r(t+j)|x(t); \tilde{\xi}) \approx f(r(t+j)|x(t); \hat{\Psi}), \quad j \geq 0\,, \quad (89)$$

where $\tilde{\xi}$ is the QML estimate of the auxiliary model coefficients obtained by fitting the model on simulated data. This approach yields filtered estimates and forecasts for the conditional mean and variance of the return via

$$\mathrm{E}\left[r(t+j)|x(t); \hat{\Psi}\right] = \int y g_K(y|x(t); \tilde{\xi})\mathrm{d}y\,, \quad (90)$$

$$\mathrm{Var}\left[r(t+j)|x(t); \hat{\Psi}\right] = \int \left(y - \mathrm{E}\left[r(t+j)|x(t); \hat{\Psi}\right]\right)^2 \times g_K(y|x(t); \tilde{\xi})\mathrm{d}y\,. \quad (91)$$

An alternative approach consists in fitting an auxiliary model for the latent variable (e. g., the return conditional variance) as a function of current and lagged returns. It is straightforward to estimate such model using data on the latent variable and the associated returns simulated from the structural model with the EMM coefficient $\hat{\Psi}$. Also in this case the auxiliary model density approximates the true one, i. e.,

$$g_K^V(V(t+j)|x(t); \tilde{\xi}) \approx f^V(V(t+j)|x(t); \hat{\psi})\,, \quad j \geq 0. \quad (92)$$

This approach yields a forecast for the conditional variance process,

$$\mathrm{E}\left[V(t+j)|x(t); \hat{\Psi}\right] = \int v g_K^V(v|x(t); \tilde{\xi})\mathrm{d}v\,. \quad (93)$$

In sum, reprojection is a simulation approach to implement a non-linear Kalman-filter-type technique, which yields effective forecasts for the unobservable state vector.

The indirect inference estimator by Gouriéroux et al. [162] and Smith [243] is closely related to the EMM estimator. Indirect inference exploits that the following two quantities should be close when the structural model is correct and the data are simulated at the true parameter Ψ_0: (i) the QML estimator $\hat{\xi}$ for the auxiliary model computed from actual data; (ii) the QML estimator $\hat{\xi}(\Psi)$ for the auxiliary model fitted on simulations from the structural model. Minimizing the distance between $\hat{\xi}$ and $\hat{\xi}(\Psi)$ in an appropriate metric yields the indirect inference estimator for Ψ. Similar to EMM, asymptotic normality holds and a χ^2 test for over-identifying restrictions is available. However, the indirect inference approach is computationally more demanding, because finding the value of Ψ that minimizes the distance function requires re-estimating the auxiliary model on a different simulated sample for each iteration of the optimization routine. EMM does not have this drawback, since the EMM objective function is evaluated at the same fitted score at each iteration. Nonetheless, there may well be circumstances where particular auxiliary models are of primary economic interest and estimation based on the corresponding moment conditions may serve as a useful diagnostic tool for model performance in such directions.

Several studies have used EMM to fit continuous-time SV jump-diffusion models for equity index returns, e. g., Andersen et al. [15], Benzoni [59], Chernov and Ghysels [93], and Chernov et al. [94,95]. Andersen and Lund [28] and Andersen et al. [16] use EMM to estimate SV jump-diffusion models for the short-term interest rate. Ahn et al. [1,2], Brandt and Chapman [71], and Dai and Singleton [114] fit different flavors of multi-factor dynamic term structure models. Andersen et al. [27] document the small-sample properties of the efficient method of moments estimator for stationary processes, while Duffee and Stanton [128] study its properties for near unit-root processes. A. Ronald Gallant and George E. Tauchen at Duke University have prepared well-documented general-purpose EMM and SNP packages, available for download at the web address ftp.econ.duke.edu in the directories pub/get/emm and pub/get/snp. In applications it is often useful to customize the SNP density to allow for a more parsimonious fit of the data under investigation. For instance, Andersen et al. [15,16], Andersen and Lund [28], and Benzoni [59] rely on the SNP density (80)–(84).

Markov Chain Monte Carlo The MCMC method provides a Bayesian solution to the inference problem for a dynamic asset pricing model. The approach treats the model coefficient Ψ as well as the vector of latent state variables X as random variables and computes the posterior distribution $f(\Psi, X|Y)$, conditional on certain observable variables Y, predicted by the model. The setting is sufficiently general to deal with a wide range of situations. For instance, X and Y can be the (latent) volatility and (observable) return processes as is the case of an SV model for asset returns. Or X and Y can be the latent state vector and observable yields in a dynamic term structure model.

The posterior distribution $f(\Psi, X|Y)$ is the main tool to draw inference not only on the coefficient Ψ but also on the latent vector X. Since $f(\Psi, X|Y)$ is unknown in closed-form in relevant applications, MCMC relies on a simulation (a Markov Chain) from the conditional density $f(\Psi, X|Y)$ to compute mode, mean, and standard deviations for the model coefficients and state variables via the Monte Carlo method.

The posterior $f(\Psi, X|Y)$ is analytically untractable and extremely high-dimensional, so that simulation directly from $f(\Psi, X|Y)$ is typically infeasible. The MCMC approach hinges on the Clifford–Hammersley theorem, which determines conditions under which the posterior $f(\Psi, X|Y)$ is uniquely determined by the marginal posterior distributions $f(\Psi|X, Y)$ and $f(X|\Psi, Y)$. In turn, the posteriors $f(\Psi|X, Y)$ and $f(X|\Psi, Y)$ are determined by a set or univariate posterior distributions. Specifically, denote with $\Psi(i)$ the ith element of the coefficient Ψ, $i = 1, \ldots, K$, and with $\Psi(-i)$ the vector consisting of all elements in Ψ except for the ith one. Similarly denote with $X(t)$ the tth row of the state vector, $t = 1, \ldots, T$, and with $X(-t)$ the rest of the vector. Then the Clifford–Hammersley theorem allows to characterize the posterior $f(\Psi, X|Y)$ via $K + T$ univariate posteriors,

$$f(\Psi(i)|\Psi(-i), X, Y) \, , \quad i = 1, \ldots, K \tag{94}$$

$$f(X(t)|X(-t), \Psi, Y) \, , \quad t = 1, \ldots, T \, . \tag{95}$$

The construction of the Markov Chain relies on the so-called Gibbs sampler. The first step of the algorithm consists in choosing initial values for the coefficient and the state, Ψ_0 and X_0. When (one of or both) the multi-dimensional posteriors are tractable, the Gibbs sampler generates values Ψ_1 and X_1 directly from $f(\Psi|X, Y)$ and $f(X|\Psi, Y)$. Alternatively, each element of Ψ_1 and X_1 is drawn from the univariate posteriors (94)–(95). Some of these posteriors may also be analytically intractable or efficient algorithms to draw from these posteriors may not exist. In such cases the Metropolis-Hastings algorithm ensures that the simulated sample is consistent with the posterior target distribution. Metropolis-Hastings sampling consists of an accept-reject procedure of the draws from a 'proposal' or 'candidate' tractable density, which is used to approximate the unknown posterior (see, e. g., Johannes and Polson [187]).

Subsequent iterations of Gibbs sampling, possibly in combination with the Metropolis-Hastings sampling, yield a series of 'sweeps' $\{\Psi_s, X_s\}$, $s = 1, \ldots, S$, with limiting distribution $f(\Psi, X|Y)$. A long number of sweeps may be necessary to 'span' the whole posterior distribution and obtain convergence due to the serial dependence of subsequent draws of coefficients and state variables. When the algorithm has converged, additional simulations provide a sample from the joint posterior distribution.

The MCMC approach has several advantages. First, the inference automatically accounts for parameter uncertainty. Further, the Markov Chain provides a direct and elegant solution to the *smoothing* problem, i. e., the problem of determining the posterior distribution for the state vector X conditional on the entire data sample, $f(X(t)|Y(1), \ldots, Y(T), \Psi)$, $t = 1, \ldots, T$. The limitation on the approach is largely that efficient sampling schemes for the posterior distribution must be constructed for each specific problem at hand which by nature is case specific and potentially cumbersome or inefficient. Nonetheless, following the development of more general simulation algorithms, the method has proven flexible for efficient estimation of a broad class of important models.

One drawback is that MCMC does not deliver an immediate solution to the *filtering* problem, i. e., determining $f(X(t)|Y(1), \ldots, Y(t), \Psi)$, and the *forecasting* problem, i. e., determining $f(X(t+j)|Y(1), \ldots, Y(t), \Psi)$, $j > 0$. However, recent research is overcoming this limitation through the use of the 'particle filter'. Bayes rule implies

$$f(X(t+1)|Y(1),\dots,Y(t+1),\Psi) \propto f(Y(t+1)| X(t+1),\Psi) f(X(t+1)|Y(1),\dots,Y(t),\Psi)\,, \quad (96)$$

where the symbol $\propto$ denotes 'proportional to'. The first density on the right-hand side of Eq. (96) is determined by the SV model and it is often known in closed form. In contrast, the second density at the far-right end of the equation is given by an integral that involves the unknown filtering density at the prior period, $f(X(t)|Y(1),\dots,Y(t),\Psi)$:

$$f(X(t{+}1)|Y(1),\dots,Y(t),\Psi) = \int f(X(t{+}1)|X(t),\Psi) \times f(X(t)|Y(1),\dots,Y(t),\Psi)\mathrm{d}X(t)\,. \quad (97)$$

The particle method relies on simulations to construct a finite set of weights $w^i(t)$ and particles $X^i(t)$, $i = 1,\dots,N$, that approximate the unknown density with a finite sum,

$$f(X(t)|Y(1),\dots,Y(t),\Psi) \approx \sum_{i=1}^{N} w^i(t)\delta_{X^i(t)}\,, \quad (98)$$

where the Dirac function $\delta_{X^i(t)}$ assigns mass one to the particle $X^i(t)$. Once the set of weights and particles are determined, it is possible to re-sample from the discretized distribution. This step yields a simulated sample $\{X^s(t)\}_{s=1}^{S}$ which can be used to evaluate the density in Eq. (97) via Monte Carlo integration:

$$f(X(t+1)|Y(1),\dots,Y(t),\Psi) \approx \frac{1}{S}\sum_{s=1}^{S} f(X(t+1)|X^s(t),\Psi)\,. \quad (99)$$

Equation (99) solves the forecasting problem while combining formulas (96) and (99) solves the filtering problem. The challenge in practical application of the particle filter is to identify an accurate and efficient algorithm to construct the set of particles and weights. We point the interested reader to Kim et al. [193], Pitt and Shephard [228] and Johannes and Polson [187] for a discussion on how to approach this problem.

The usefulness of the MCMC method to solve the inference problem for SV models has been evident since the early work by Jacquier et al. [180], who develop an MCMC algorithm for the logarithmic SV model. Jacquier et al. [181] provide extensions to correlated and non-normal error distributions. Kim et al. [193], Pitt and Shephard [228] and Chib et al. [96] develop simulation-based methods to solve the filtering problem, while Chib et al. [97] use the MCMC approach to estimate a multivariate SV model. Elerian et al. [135] and Eraker [140] discuss how to extend the MCMC inference method to a continuous-time setting. Eraker [140] uses the MCMC approach to estimate an SV diffusion process for interest rates, while Jones [189] estimates a continuous-time model for the spot rate with non-linear drift function. Eraker et al. [142] estimate an SV jump-diffusion process using data on S&P 500 return while Eraker [141] estimates a similar model using joint data on options and underlying S&P 500 returns. Li et al. [196] allow for Lévy-type jumps in their model. Collin-Dufresne et al. [104] use the MCMC approach to estimate multi-factor affine dynamic term structure model using swap rates data. Johannes and Polson [186] give a comprehensive survey of the still ongoing research on the use of the MCMC approach in the general nonlinear jump-diffusion SV setting.

Estimation from Option Data

Options' payoffs are non-linear functions of the underlying security price. This feature renders options highly sensitive to jumps in the underlying price and to return volatility, which makes option data particularly useful to identify return dynamics. As such, several studies have taken advantage of the information contained in option prices, possibly in combination with underlying return data, to estimate SV models with or without discontinuities in returns and volatility.

Applications to derivatives data typically require a model for the pricing errors. A common approach is to posit that the market price of an option, O^*, normalized by the underlying observed security price S^*, is the sum of the normalized model-implied option price, O/S^*, and a disturbance term ε (e. g., Renault [230]):

$$\frac{O^*}{S^*} = \frac{O(S^*,V,K,\tau,\Psi)}{S^*} + \varepsilon\,, \quad (100)$$

where V is the latent volatility state, K is the option strike price, τ is time to maturity, and Ψ is the vector with the model coefficients. A pricing error ε could arise for several reasons, including measurement error (e. g., price discreteness), asynchroneity between the derivatives and underlying price observations, microstructure effects, and perhaps most importantly specification error. The structure imposed on ε depends on the choice of a specific 'loss function' used for estimation (e. g., Christoffersen and Jacobs [98]). Several studies have estimated the coefficient vector Ψ by minimizing the sum of the squared option pricing errors normalized by the underlying price S^*, as in Eq. (100). Others have focused on either squared dollar pricing errors, or squared errors normalized by the options market price (instead of S^*). The latter approach has the advantage that a $1 error on an expensive in-the-money option carries less weight than the same error on a cheaper out-of-the-money contract. The drawback is that giving a lot of weight to the pricing errors on short-maturity deep-out-of-the-money options could bias the estimation results. Finally, the common practice of expressing option prices in terms of their Black-Scholes implied volatilities has inspired other scholars to minimize the deviations between Black-Scholes implied volatilities inferred from model and market prices (e. g., Mizrach [216]). An alternative course is to form a moment-based loss function and follow a GMM- or SMM-type approach to estimate Ψ. To this end moment conditions stem from distributional assumptions on the pricing error ε (e. g., $\mathrm{E}[\varepsilon] = 0$) or from the scores of a reduced-form model that approximates the data.

In estimating the model, some researchers have opted to use a panel of options consisting of contracts with multiple strikes and maturities across dates in the sample period. This choice brings a wealth of information on the cross-sectional and term-structure properties of the implied volatility smirk into the analysis. Others rely on only one option price observation per time period, which shifts the focus to the time-series dimension of the data. Some studies re-estimate the model on a daily basis rather than seeking a single point estimate for the coefficient Ψ across the entire sample period. This ad hoc approach produces smaller in-sample pricing errors, which can be useful to practitioners, but at the cost of concealing specification flaws by over-fitting the model, which tends to hurt out-of-sample performance. The different approaches are in part dictated by the intended use of the estimated system as practitioners often are concerned with market making and short-term hedging while academics tend to value stable relations that may form the basis for consistent modeling of the dominant features of the system over time.

Early contributions focus on loss functions based on the sum of squared option pricing errors and rely entirely on option data for estimation. This approach typically yields an estimate of the model coefficient Ψ that embeds an adjustment for risk, i. e., return and volatility dynamics are identified under the risk-neutral rather than the physical probability measure. For instance, Bates [56] considers an SV jump-diffusion model for Deutsche Mark foreign currency options and estimates its coefficient vector Ψ via nonlinear generalized least squares of the normalized pricing errors with daily option data from January 1984 to June 1991. A similar approach is followed by Bates [57] who fits an SV model with two latent volatility factors and jumps using daily data on options on the S&P 500 futures from January 1988 to December 1993. Bakshi et al. [34] focus on the pricing and hedging of daily S&P 500 index options from June 1988 to May 1991. In their application they re-calibrate the model on a daily basis by minimizing the sum of the squared dollar pricing errors across options with different maturities and strikes. Huang and Wu [173] explore the pricing implications of the time-changed Lévy process by Carr and Wu [84] for daily S&P 500 index options from April 1999 to May 2000. Their Lévy return process allows for discontinuities that exhibit higher jump frequencies compared to the finite-intensity Poisson jump processes in Equations (37)–(41). Further, their model allows for a random time change, i. e., a monotonic transformation of the time variable which generates SV in the diffusion and jump components of returns. In contrast, Bakshi et al. [35] fit an SV jump-diffusion model by SMM using daily data on long-maturity S&P 500 options (LEAPS).

More recent studies have relied on joint data on S&P 500 option prices and underlying index returns, spanning different periods, to estimate the model. This approach forces the same model to price securities in two different markets and relies on information from the derivatives and underlying securities to better pin down model coefficients and risk premia. For instance, Eraker [141] and Jones [190] fit different flavors of the SV model (with and without jumps, respectively) by MCMC. Pan [220] follows a GMM approach to estimate an SV jump-diffusion model using weekly data. She relies on a single at-the-money option price observation each week, which identifies the level of the latent volatility state variable (i. e., at each date she fixes the error term ε at zero and solves Eq. (100) for V). Aït-Sahalia and Kimmel [7] apply Aït-Sahalia's [4] method to approximate the likelihood function for a joint sample of options and underlying prices. Chernov and Ghysels [93] and Benzoni [59] obtain moment conditions from the scores of a SNP auxiliary model. Similarly, other recent studies have found it useful to use joint derivatives and interest rate data to fit dynamic term structure models, e. g., Almeida et al. [9], and Bikbov and Chernov [62].

Finally, a different literature has studied the option pricing implications of a model in which asset return volatility is a deterministic function of the asset price and time, e. g., Derman and Kani [119], Dupire [134], Rubinstein [233], and Jackwerth and Rubinstein [177]. Since volatility is not stochastic in this setting, we do not review these models here and point the interested reader to, e. g., [133] for an empirical analysis of their performance.

Future Directions

In spite of much progress in our understanding of volatility new challenges lie ahead. In recent years a wide array of volatility-sensitive products has been introduced. The market for these derivatives has rapidly grown in size and complexity. Research faces the challenge to price and hedge these new products. Moreover, the recent developments in model-free volatility modeling have effectively given empirical content to the latent volatility variable, which opens the way for a new class of estimation methods and specification tests for SV systems. Related, improved volatility measures enable us to shed new light on the properties and implications of the volatility risk premium. Finally, more work is needed to better understand the linkage between fluctuations in economic fundamentals and low- and high-frequency volatility movements. We conclude this chapter by briefly reviewing some open issues in these four areas of research.

Volatility and Financial Markets Innovation

Volatility is a fundamental input to any financial and real investment decision. Markets have responded to investors' needs by offering an array of volatility-linked instruments. In 1993 the Chicago Board Option Exchange (CBOE) has introduced the VIX index, which measures the market expectations of near-term volatility conveyed by equity-index options. The index was originally computed using the Black-Scholes implied volatilities of eight different S&P 100 option (OEX) series so that, at any given time, it represented the implied volatility of a hypothetical at-the-money OEX option with exactly 30 days to expiration (see [257]). On September 22, 2003, the CBOE began disseminating price level information using a revised 'model-free' method for the VIX index. The new VIX is given by the price of a portfolio of S&P 500 index options and incorporates information from the volatility smirk by using a wider range of strike prices rather than just at-the-money series (see [77]). On March 26, 2004, trading in futures on the VIX Index started on the CBOE Futures Exchange (CFE) while on February 24, 2006, options on the VIX began trading on the Chicago Board Options Exchange. These developments have opened the way for investors to trade on option-implied measures of market volatility. The popularity of the VIX prompted the CBOE to introduce similar indices for other markets, e. g., the VXN NASDAQ 100 Volatility Index.

Along the way, a new over-the-counter market for volatility derivatives has also rapidly grown in size and liquidity. Volatility derivatives are contracts whose payments are expressed as functions of realized variance. Popular examples are variance swaps, which at maturity pay the difference between realized variance and a fixed strike price. According to estimates by BNP Paribas reported by the Risk [176] magazine, the daily trading volume for variance swaps on indices reached $4–5 million in vega notional (measured in dollars per volatility point) in 2006, which corresponds to payments in excess of $1 billion per percentage point of volatility on an annual basis (Carr and Lee [82]). Using variance swaps hedge fund managers and proprietary traders can easily place huge bets on market volatility.

Finally, in recent years credit derivatives markets have evolved in complexity and grown in size. Among the most popular credit derivatives are the credit default swaps (CDS), which provide insurance against the risk of default by a particular company. The buyer of a single-name CDS acquires the right to sell bonds issued by the company at face value when a credit event occurs. Multiple-name contracts can be purchased simultaneously through credit indices. For instance, the CDX indices track the credit spreads for different portfolios of North American companies while the iTraxx Europe indices track the spreads for portfolios of European companies. At the end of 2006 the notional amount of outstanding over-the-counter single- and multi-name CDS contracts stood at $19 and $10 trillion, respectively, according to the September 2007 Bank for International Settlements Quarterly Review.

These market developments have raised new interesting issues for research to tackle. The VIX computations based on the new model-free definition of implied volatility used by the CBOE requires the use of options with strike prices that cover the entire support of the return distribution. In practice, liquid options satisfying this requirement often do not exist and the CBOE implementation introduces random noise and systematic error into the index (Jiang and Tian [185]). Related, the VIX implementation entails a truncation, i. e., the CBOE discards illiquid option prices with strikes lying in the tails of the return distribution. As such, the notion of the VIX is more directly linked to that of corridor volatility [26]. In sum, robust implementation of a model free measure of implied volatility is still an open area of research. Future developments in this direction will also have important repercussions on the hedging practices for implied-volatility derivatives.

Pricing and hedging of variance derivatives is another active area of research. Variance swaps admit a simple replication strategy via static positions in call and put options on the underlying asset, similar to model-free implied volatility measures (e. g., [77,83]). In contrast, it is still an open area of research to determine the replication strategy for derivatives whose payoffs are non-linear function of realized variance, e. g., volatility swaps, which pay the square-root of realized variance, or call and put options on realized variance. [82] is an interesting paper in this direction.

Limited liability gives shareholders the option to default on the firm's debt obligation. As such, a debt claim has features similar to a short position in a put option. The pricing of corporate debt is therefore sensitive to the volatility of the firms' assets: higher volatility increases the probability of default and therefore reduces the price of debt and increases credit spreads. The insights and techniques developed in the SV literature could prove useful in credit risk modeling and applications (e. g., [179,248,260]).

The Use of Realized Volatility for Estimation of SV Models

Another promising line of research aims at extracting the information in RV measures for the estimation of dynamic asset pricing models. Early work along these lines includes Barndorff-Nielson and Shephard [51], who decompose RV into actual volatility and realized volatility error. They consider a state-space representation for the decomposition and apply the Kalmann filter to estimate different flavors of the SV model. Moreover, Bollerslev and Zhou [68] and Garcia et al. [156], build on the insights of Meddahi [210] to estimate SV diffusion models using conditional moments of integrated volatility. More recently, Todorov [252] generalizes the analysis for the presence of jumps.

Related, recent studies have started to use RV measures to test the implications of models previously estimated with lower-frequency data. Since RV gives empirical content to the latent quadratic variation process, this approach allows for a direct test of the model-implied restrictions on the latent volatility factor. Recent work along these lines includes Andersen and Benzoni [12], who use model-free RV measures to show that the volatility spanning condition embedded in some affine term structure models is violated in the US Treasury market. Christoffersen et al. [100] note that the Heston square-root SV model implies that the dynamics for the standard deviation process are conditionally Gaussian. They reject this condition by examining the distribution of the changes in the square-root RV measure for S&P 500 returns.

Volatility Risk Premium

More work is needed to better understand the link between asset return volatility and model risk premia. Also in this case, RV measures are a fruitful source of information to shed new light on the issue. Among the recent studies that pursue this venue is Bollerslev et al. [66], who exploit the moments of RV and option-implied volatility to gauge a measure of the volatility risk premium. Todorov [251] explores the variance risk premium dynamics using high-frequency S&P 500 index futures data and data on the VIX index. He finds the variance risk premium to vary significantly over time and to increase during periods of high volatility and immediately after big jumps in underlying returns. Carr and Wu [85] provide a broader analysis of the variance risk premium for five equity indices and 35 individual stocks. They find the premium to be large and negative for the indices while it is much smaller for the individual stocks. Further, they also find the premium to increase (in absolute value) with the level of volatility. Additional work on the volatility risk premium embedded in individual stock options is in Bakshi and Kapadia [36], Driessen et al. [123], and Duarte and Jones [126]. Other studies have examined the linkage between volatility risk premia and equity returns (e. g., [69]) and hedge-fund performance (e. g., [70]). New research is also examining the pricing of aggregate volatility risk in the cross-section of stock returns. For instance, Ang et al. [30] find that average returns are lower on stocks that have high sensitivities to innovations in aggregate volatility and high idiosyncratic volatility (see also the related work by Chen [90] Ang et al. [32], Bandi et al, Guo et al. [42]). This evidence is consistent with the findings of the empirical option pricing literature, which suggests that there is a negative risk premium for volatility risk. Intuitively, periods of high market volatility are associated to worsened investment opportunities and tend to coincide with negative stock market returns (the so-called leverage effect). As such, investors are willing to pay higher prices (i. e., accept lower expected returns) to hold stocks that do well in high-volatility conditions.

Determinants of Volatility

Finally, an important area of future research concerns the linkage between asset return volatility and economic uncertainty. Recent studies have proposed general equilibrium models that produce low-frequency fluctuations in conditional volatility, e. g., Campbell and Cochrane [80], Bansal and Yaron [47], McQueen and Vorkink [207], and Tauchen [246]. Related, Engle and Rangel [139] and Engle et al. [138] link macroeconomic variables and long-run volatility movements. It is still an open issue, however, to determine the process through which news about economic fundamentals are embedded into prices to generate high-frequency volatility fluctuations. Early research by Schwert [236] and Shiller [241] has concluded that the amplitude of the fluctuations in aggregate stock volatility is difficult to explain using simple models of stock valuation. Further, Schwert [236] notes that while aggregate leverage is significantly correlated with volatility, it explains a relatively small part of the movements in stock volatility. Moreover, he finds little evidence that macroeconomic volatility (measured by inflation and industrial production volatility) helps predict future asset return volatility. Model-free realized volatility measures are a useful tool to further investigate this issue. Recent work in this direction includes Andersen et al. [22] and Andersen and Bollerslev [17], who explore the linkage between news arrivals and exchange rates volatility, and Andersen and Benzoni [13], who investigate the determinants of bond yields volatility in the US Treasury market. Related, Balduzzi et al. [38] and Fleming and Remolona [146] study the reaction of trading volume, bid-ask spread, and price volatility to macroeconomic surprises in the US Treasury market, while Brandt and Kavajecz [74] and Pasquariello and Vega [222] focus instead on the price discovery process and explore the implications of order flow imbalances (excess buying or selling pressure) on day-to-day variation in yields.

Acknowledgments

We are grateful to Olena Chyruk, Bruce Mizrach (the Section Editor) and Neil Shephard for helpful comments and suggestions. Of course, all errors remain our sole responsibility. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Chicago or the Federal Reserve System. The work of Andersen is supported by a grant from the NSF to the NBER and support from CREATES funded by the Danish National Research Foundation.

Bibliography

Primary Literature

1. Ahn DH, Dittmar RF, Gallant AR (2002) Quadratic Term Structure Models: Theory and Evidence. Rev Finance Stud 15:243–288
2. Ahn DH, Dittmar RF, Gallant AR, Gao B (2003) Purebred or hybrid?: Reproducing the volatility in term structure dynamics. J Econometrics 116:147–180
3. Aït-Sahalia Y (2002) Maximum-Likelihood Estimation of Discretely-Sampled Diffusions: A Closed-Form Approximation Approach. Econometrica 70:223–262
4. Aït-Sahalia Y (2007) Closed-Form Likelihood Expansions for Multivariate Diffusions. Annals of Statistics, forthcoming
5. Aït-Sahalia Y, Hansen LP, Scheinkman J (2004) Operator Methods for Continuous-Time Markov Processes. In: Hansen LP, Aït-Sahalia Y (eds) Handbook of Financial Econometrics. North-Holland, Amsterdam, forthcoming
6. Aït-Sahalia Y, Jacod J (2007) Testing for Jumps in a Discretely Observed Process. Annals of Statistics, forthcoming
7. Aït-Sahalia Y, Kimmel R (2007) Maximum Likelihood Estimation of Stochastic Volatility Models. J Finance Econ 83:413–452
8. Alizadeh S, Brandt MW, Diebold FX (2002) Range-Based Estimation of Stochastic Volatility Models. J Finance 57:1047–1091
9. Almeida CIR, Graveline JJ, Joslin S (2006) Do Options Contain Information About Excess Bond Returns? Working Paper, UMN, Fundação Getulio Vargas, MIT, Cambridge
10. Andersen TG (1994) Stochastic Autoregressive Volatility: A Framework for Volatility Modeling. Math Finance 4:75–102
11. Andersen TG (1996) Return Volatility and Trading Volume: An Information Flow Interpretation of Stochastic Volatility. J Finance 51:169–204
12. Andersen TG, Benzoni L (2006) Do Bonds Span Volatility Risk in the US Treasury Market? A Specification Test for Affine Term Structure Models. Working Paper, KSM and Chicago FED, Chicago
13. Andersen TG, Benzoni L (2007) The Determinants of Volatility in the US Treasury market. Working Paper, KSM and Chicago FED, Chicago
14. Andersen TG, Benzoni L (2007) Realized volatility. In: Andersen TG, Davis RA, Kreiss JP, Mikosch T (eds) Handbook of Financial Time Series. Springer, Berlin (forthcoming)
15. Andersen TG, Benzoni L, Lund J (2002) An Empirical Investigation of Continuous-Time Equity Return Models. J Finance 57:1239–1284
16. Andersen TG, Benzoni L, Lund J (2004) Stochastic Volatility, Mean Drift and Jumps in the Short Term Interest Rate. Working Paper, Northwestern University, University of Minnesota, and Nykredit Bank, Copenhagen
17. Andersen TG, Bollerslev T (1998) Deutsche Mark-Dollar Volatility: Intraday Activity Patterns, Macroeconomic Announcements, and Longer Run Dependencies. J Finance 53:219–265
18. Andersen TG, Bollerslev T (1998) Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts. Int Econ Rev 39:885–905
19. Andersen TG, Bollerslev T, Diebold FX (2004) Parametric and nonparametric volatility measurement. In: Hansen LP, Aït-Sahalia Y (eds) Handbook of Financial Econometrics. North-Holland, Amsterdam, forthcoming
20. Andersen TG, Bollerslev T, Diebold FX (2007) Roughing It Up: Including Jump Components in Measuring, Modeling and Forecasting Asset Return Volatility. Rev Econ Statist 89:701–720
21. Andersen TG, Bollerslev T, Diebold FX, Labys P (2003) Modeling and Forecasting Realized Volatility. Econometrica 71:579–625
22. Andersen TG, Bollerslev T, Diebold FX, Vega C (2003) Micro Effects of Macro Announcements: Real-Time Price Discovery in Foreign Exchange. Ammer Econom Rev 93:38–62
23. Andersen TG, Bollerslev T, Dobrev D (2007) No-arbitrage semi-martingale restrictions for continuous-time volatility models subject to leverage effects, jumps and i.i.d. noise: Theory and testable distributional implications. J Econome 138:125–180

24. Andersen TG, Bollerslev T, Meddahi N (2004) Analytic Evaluation of Volatility Forecasts. Int Econ Rev 45:1079–1110
25. Andersen TG, Bollerslev T, Meddahi N (2005) Correcting the Errors: Volatility Forecast Evaluation Using High-Frequency Data and Realized Volatilities. Econometrica 73:279–296
26. Andersen TG, Bondarenko O (2007) Construction and Interpretation of Model-Free Implied Volatility. Working Paper, KSM and UIC, Chicago
27. Andersen TG, Chung HJ, Sørensen BE (1999) Efficient Method of Moments Estimation of a Stochastic Volatility Model: A Monte Carlo Study. J Econom 91:61–87
28. Andersen TG, Lund J (1997) Estimating continuous-time stochastic volatility models of the short term interest rate diffusion. J Econom 77:343–377
29. Ané T, Geman H (2000) Order Flow, Transaction Clock, and Normality of Asset Returns. J Finance 55:2259–2284
30. Ang A, Hodrick RJ, Xing Y, Zhang X (2006) The Cross-Section of Volatility and Expected Returns. J Finance 51:259–299
31. Ang A, Hodrick RJ, Xing Y, Zhang X (2008) High idiosyncratic volatility and low returns: international and further U.S. evidence. Finance Econ (forthcoming)
32. Bachelier L (1900) Théorie de la Spéculation. Annales de École Normale Supérieure 3, Gauthier-Villars, Paris. English translation: Cootner PH (ed) (1964) The Random Character of Stock Market Prices. MIT Press, Cambridge
33. Back K (1991) Asset Prices for General Processes. J Math Econ 20:371–395
34. Bakshi G, Cao C, Chen Z (1997) Empirical Performance of Alternative Option Pricing Models. J Finance 52:2003–2049
35. Bakshi G, Cao C, Chen Z (2002) Pricing and hedging long-term options. J Econom 94:277–318
36. Bakshi G, Kapadia N (2003) Delta-Hedged Gains and the Negative Market Volatility Risk Premium. Rev Finance Stud 16:527–566
37. Bakshi G, Kapadia N, Madan D (2003) Stock Return Characteristics, Skew Laws, and the Differential Pricing of Individual Equity Options. Rev Finance Stud 16:101–143
38. Balduzzi P, Elton EJ, Green TC (2001) Economic News and Bond Prices: Evidence from the US Treasury Market. J Finance Quant Anal 36:523–543
39. Bandi FM (2002) Short-term interest rate dynamics: a spatial approach. J Financ Econ 65:73–110
40. Bandi FM, Moise CE, Russel JR (2008) Market volatility, market frictions, and the cross section of stock returns. Working Paper, University of Chicago, and Case Western Reverse University, Cleveland
41. Bandi FM, Nguyen T (2003) On the functional estimation of jump-diffusion models. J Econom 116:293–328
42. Bandi FM, Phillips PCB (2002) Nonstationary Continuous-Time Processes. In: Hansen LP, Aït-Sahalia Y (eds) Handbook of Financial Econometrics. North-Holland, Amsterdam, forthcoming
43. Bandi FM, Phillips PCB (2003) Fully nonparametric estimation of scalar diffusion models. Econometrica 71:241–283
44. Bandi FM, Phillips PCB (2007) A simple approach to the parametric estimation of potentially nonstationary diffusions. J Econom 137:354–395
45. Bandi FM, Russell J (2006) Separating Microstructure Noise from Volatility. J Finance Econ 79:655–692
46. Bandi FM, Russell J (2007) Volatility. In: Birge J, Linetsky V (eds) Handbook of Financial Engineering. Elsevier, Amsterdam
47. Bansal R, Yaron A (2004) Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles. J Finance 59:1481–1509
48. Bartlett MS (1938) The Characteristic Function of a Conditional Statistic. J Lond Math Soc 13:63–67
49. Barndorff-Nielsen OE, Hansen P, Lunde A, Shephard N (2007) Designing Realized Kernels to Measure the Ex-Post Variation of Equity Prices in the Presence of Noise. Working Paper, University of Aarhus, Aarhus. Stanford University, Nuffield College, Oxford
50. Barndorff-Nielsen OE, Shephard N (2001) Non-Gaussian Ornstein-Uhlenbeckbased models and some of their uses in financial economics. J R Stat Soc B 63:167–241
51. Barndorff-Nielsen OE, Shephard N (2002) Econometric Analysis of Realised Volatility and its Use in Estimating Stochastic Volatility Models. J R Stat Soc B 64:253–280
52. Barndorff-Nielsen OE, Shephard N (2002b) Estimating quadratic variation using realized variance. J Appl Econom 17:457–477
53. Barndorff-Nielsen OE, Shephard N (2004) Power and bipower variation with stochastic volatility and jumps. J Finance Econom 2:1–37
54. Barndorff-Nielsen OE, Shephard N (2006) Econometrics of testing for jumps in financial economics using bipower variation. J Finance Econom 4:1–30
55. Bates DS (1991) The Crash of ’87: Was It Expected? The Evidence from Options Markets. J Finance 46:1009–1044
56. Bates DS (1996) Jumps and stochastic volatility: exchange rate processes implicit in deutsche mark options. Rev Finance Stud 9:69–107
57. Bates DS (2000) Post-’87 crash fears in the S&P 500 futures option market. J Econom 94:181–238
58. Bates DS (2006) Maximum Likelihood Estimation of Latent Affine Processes. Rev Finance Stud 19:909–965
59. Benzoni L (2002) Pricing Options under Stochastic Volatility: An Empirical Investigation. Working Paper, Chicago FED
60. Benzoni L, Collin-Dufresne P, Goldstein RS (2007) Explaining Pre- and Post-1987 Crash Prices of Equity and Options within a Unified General Equilibrium Framework. Working Paper, Chicago FED, UCB, and UMN, Minneapolis
61. Bibby BM, Jacobsen M, Sorensen M (2004) Estimating Functions for Discretely Sampled Diffusion-Type Models. In: Hansen LP, Aït-Sahalia Y (eds) Handbook of Financial Econometrics. North-Holland, Amsterdam, forthcoming
62. Bikbov R, Chernov M (2005) Term Structure and Volatility: Lessons from the Eurodollar Markets. Working Paper, LBS, Deutsche Bank, New York
63. Black F, Scholes M (1973) The Pricing of Options and Corporate Liabilities. J Political Econ 81:637–654
64. Bladt M, Sørensen M (2007) Simple Simulation of Diffusion Bridges with Application to Likelihood Inference for Diffusions. Working Paper, University of Copenhagen, Copenhagen
65. Bollen NPB, Whaley RE (2004) Does Net Buying Pressure Affect the Shape of Implied Volatility Functions? J Finance 59:711–753
66. Bollerslev T, Gibson M, Zhou H (2004) Dynamic Estimation of Volatility Risk Premia and Investor Risk Aversion from Option-Implied and Realized Volatilities. Working Paper, Duke University, Federal Reserve Board, Washington D.C.

67. Bollerslev T, Jubinsky PD (1999) Equity Trading vol and Volatility: Latent Information Arrivals and Common Long-Run Dependencies. J Bus Econ Stat 17:9–21
68. Bollerslev T, Zhou H (2002) Estimating stochastic volatility diffusion using conditional moments of integrated volatility. J Econom 109:33–65
69. Bollerslev T, Zhou H (2007) Expected Stock Returns and Variance Risk Premia. Working Paper, Duke University and Federal Reserve Board, Washington D.C.
70. Bondarenko O (2004) Market price of variance risk and performance of hedge funds. Working Paper, UIC, Chicago
71. Brandt MW, Chapman DA (2003) Comparing Multifactor Models of the Term Structure. Working Paper, Duke University and Boston College, Chestnut Hill
72. Brandt MW, Diebold FX (2006) A No-Arbitrage Approach to Range-Based Estimation of Return Covariances and Correlations. J Bus 79:61–73
73. Brandt MW, Jones CS (2006) Volatility Forecasting with Range-Based EGARCH Models. J Bus Econ Stat 24:470–486
74. Brandt MW, Kavajecz KA (2004) Price Discovery in the US Treasury Market: The Impact of Orderflow and Liquidity on the Yield Curve. J Finance 59:2623–2654
75. Brandt MW, Santa-Clara P (2002) Simulated likelihood estimation of diffusions with an application to exchange rate dynamics in incomplete markets. J Finance Econ 63:161–210
76. Breidt FJ, Crato N, de Lima P (1998) The detection and estimation of long memory in stochastic volatility. J Econom 83:325–348
77. Britten-Jones M, Neuberger A (2000) Option Prices, Implied Price Processes, and Stochastic Volatility. J Finance 55:839–866
78. Broadie M, Chernov M, Johannes MJ (2007) Model Specification and Risk Premia: Evidence from Futures Options. J Finance 62:1453–1490
79. Buraschi A, Jackwerth J (2001) The price of a smile: hedging and spanning in option markets. Rev Finance Stud 14:495–527
80. Campbell JY, Cochrane JH (1999) By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior. J Polit Economy 107:205–251
81. Campbell JY, Viceira LM (2001) Who Should Buy Long-Term Bonds? Ammer Econ Rev 91:99–127
82. Carr P, Lee R (2007) Robust Replication of Volatility Derivatives. Working Paper, NYU and Uinversity of Chicago, Chicago
83. Carr P, Madan D (1998) Towards a theory of volatility trading. In: Jarrow R (ed) Volatility. Risk Publications, London
84. Carr P, Wu L (2004) Time-changed Lévy processes and option pricing. J Finance Econ 71:113–141
85. Carr P, Wu L (2007) Variance Risk Premia. Rev Finance Stud, forthcoming
86. Carrasco M, Chernov M, Florens JP, Ghysels E (2007) Efficient estimation of general dynamic models with a continuum of moment conditions. J Econom 140:529–573
87. Carrasco M, Florens JP (2000) Generalization of GMM to a continuum of moment conditions. Econom Theory 16:797–834
88. Chacko G, Viceira LM (2003) Spectral GMM estimation of continuous-time processes. J Econom 116:259–292
89. Chan KC, Karolyi GA, Longstaff FA, Sanders AB (1992) An Empirical Comparison of Alternative Models of the Short-Term Interest Rate. J Finance 47:1209–1227
90. Chen J (2003) Intertemporal CAPM and the Cross-Section of Stock Return. Working Paper, USC, Los Angeles
91. Chen RR, Scott L (1993) Maximum likelihood estimation for a multifactor equilibrium model of the term structure of interest rates. J Fixed Income 3:14–31
92. Cheridito P, Filipović D, Kimmel RL (2007) Market Price of Risk Specifications for Affine Models: Theory and Evidence. J Finance Econ, 83(1):123–170
93. Chernov M, Ghysels E (2002) A study towards a unified approach to the joint estimation of objective and risk neutral measures for the purpose of options valuation. J Finance Econ 56:407–458
94. Chernov M, Gallant AR, Ghysels E, Tauchen G (1999) A New Class of Stochastic Volatility Models with Jumps: Theory and Estimation. Working Paper, London Business School, Duke University, University of Northern Carolina
95. Chernov M, Gallant AR, Ghysels E, Tauchen G (2003) Alternative models for stock price dynamics. J Econom 116:225–257
96. Chib S, Nardari F, Shephard N (2002) Markov chain Monte Carlo methods for stochastic volatility models. J Econom 108:281–316
97. Chib S, Nardari F, Shephard N (2006) Analysis of high dimensional multivariate stochastic volatility models. J Econom 134:341–371
98. Christoffersen PF, Jacobs K (2004) The importance of the loss function in option valuation. J Finance Econ 72:291–318
99. Christoffersen PF, Jacobs K, Karoui L, Mimouni K (2007) Estimating Term Structure Models Using Swap Rates. Working Paper, McGill University, Montreal
100. Christoffersen PF, Jacobs K, Mimouni K (2006a) Models for S&P 500 Dynamics: Evidence from Realized Volatility, Daily Returns, and Option Prices. Working Paper, McGill University, Montreal
101. Christoffersen PF, Jacobs K, Wang Y (2006b) Option Valuation with Long-run and Short-run Volatility Components. Working Paper, McGill University, Montreal
102. Clark PK (1973) A Subordinated Stochastic Process Model with Finite Variance for Speculative Prices. Econometrica 41:135–156
103. Collin-Dufresne P, Goldstein RS (2002) Do Bonds Span the Fixed Income Markets? Theory and Evidence for Unspanned Stochastic Volatility. J Finance 57:1685–1730
104. Collin-Dufresne P, Goldstein R, Jones CS (2008) Identification of Maximal Affine Term Structure Models. J Finance 63(2):743–795
105. Collin-Dufresne P, Goldstein R, Jones CS (2007b) Can Interest Rate Volatility be Extracted from the Cross Section of Bond Yields? An Investigation of Unspanned Stochastic Volatility. Working Paper, UCB, USC, UMN, Minneapolis
106. Collin-Dufresne P, Solnik B (2001) On the Term Structure of Default Premia in the Swap and LIBOR Markets. J Finance 56:1095–1115
107. Comte F, Coutin L, Renault E (2003) Affine Fractional Stochastic Volatility Models with Application to Option Pricing. Working Paper, CIRANO, Montreal
108. Comte F, Renault E (1998) Long memory in continuous-time stochastic volatility models. Math Finance 8:291–323
109. Conley TG, Hansen LP, Luttmer EGJ, Scheinkman JA (1997) Short-term interest rates as subordinated diffusions. Rev Finance Stud 10:525–577

110. Corsi F (2003) A Simple Long Memory Model of realized Volatility. Working paper, University of Southern Switzerland, Lugano
111. Coval JD, Shumway T (2001) Expected Option Returns. J Finance 56:983–1009
112. Cox JC, Ingersoll JE, Ross SA (1985) An Intertemporal General Equilibrium Model of Asset Prices. Econometrica 53:363–384
113. Cox JC, Ingersoll JE, Ross SA (1985) A Theory of the Term Structure of Interest Rates. Econometrica 53:385–407
114. Dai Q, Singleton KJ (2000) Specification Analysis of Affine Term Structure Models. J Finance 55:1943–1978
115. Dai Q, Singleton KJ (2003) Term Structure Dynamics in Theory and Reality. Rev Finance Stud 16:631–678
116. Danielsson J (1994) Stochastic Volatility in Asset Prices: Estimation by Simulated Maximum Likelihood. J Econom 64:375–400
117. Danielsson J, Richard JF (1993) Accelerated Gaussian Importance Sampler with Application to Dynamic Latent Variable Models. J Appl Econom 8:S153–S173
118. Deo R, Hurvich C (2001) On the Log Periodogram Regression Estimator of the Memory Parameter in Long Memory Stochastic Volatility Models. Econom Theory 17:686–710
119. Derman E, Kani I (1994) The volatility smile and its implied tree. Quantitative Strategies Research Notes, Goldman Sachs, New York
120. Diebold FX, Nerlove M (1989) The Dynamics of Exchange Rate Volatility: A Multivariate Latent Factor ARCH Model. J Appl Econom 4:1–21
121. Diebold FX, Strasser G (2007) On the Correlation Structure of Microstructure Noise in Theory and Practice. Working Paper, University of Pennsylvania, Philadelphia
122. Dobrev D (2007) Capturing Volatility from Large Price Moves: Generalized Range Theory and Applications. Working Paper, Federal Reserve Board, Washington D.C.
123. Driessen J, Maenhout P, Vilkov G (2006) Option-Implied Correlations and the Price of Correlation Risk. Working Paper, University of Amsterdam and INSEAD, Amsterdam
124. Duarte J (2004) Evaluating An Alternative Risk Preference in Affine Term Structure Models. Rev Finance Stud 17:370–404
125. Duarte J (2007) The Causal Effect of Mortgage Refinancing on Interest-Rate Volatility: Empirical Evidence and Theoretical Implications. Rev Finance Stud, forthcoming
126. Duarte J, Jones CS (2007) The Price of Market Volatility Risk. Working Paper, University of Washington and USC, Washington
127. Duffee GR (2002) Term Premia and Interest Rate Forecasts in Affine Models. J Finance 57:405–443
128. Duffee G, Stanton R (2008) Evidence on simulation inference for near unit-root processes with implications for term structure estimation. J Finance Econom 6:108–142
129. Duffie D, Kan R (1996) A yield-factor model of interest rates. Math Finance 6:379–406
130. Duffie D, Pan J, Singleton KJ (2000) Transform Analysis and Asset Pricing for Affine Jump-Diffusions. Econometrica 68:1343–1376
131. Duffie D, Singleton KJ (1993) Simulated Moments Estimation of Markov Models of Asset Prices. Econometrica 61:929–952
132. Duffie D, Singleton KJ (1997) An Econometric Model of the Term Structure of Interest-Rate Swap Yields. J Finance 52:1287–1321
133. Dumas B, Fleming J, Whaley RE (1996) Implied Volatility Functions: Empirical Tests. J Finance 53:2059–2106
134. Dupire B (1994) Pricing with a smile. Risk 7:18–20
135. Elerian O, Chib S, Shephard N (2001) Likelihood Inference for Discretely Observed Nonlinear Diffusions. Econometrica 69:959–994
136. Engle RF (2002) New frontiers for ARCH models. J Appl Econom 17:425–446
137. Engle RF, Gallo GM (2006) A multiple indicators model for volatility using intra-daily data. J Econom 131:3–27
138. Engle RF, Ghysels E, Sohn B (2006) On the Economic Sources of Stock Market Volatility. Working Paper, NYU and UNC, Chapel Hill
139. Engle RF, Rangel JG (2006) The Spline-GARCH Model for Low Frequency Volatility and Its Global Macroeconomic Causes. Working Paper, NYU, New York
140. Eraker B (2001) MCMC Analysis of Diffusions with Applications to Finance. J Bus Econ Stat 19:177–191
141. Eraker B (2004) Do Stock Prices and Volatility Jump? Reconciling Evidence from Spot and Option Prices. J Finance 59:1367–1404
142. Eraker B, Johannes MS, Polson N (2003) The Impact of Jumps in Volatility and Returns. J Finance 58:1269–1300
143. Fan R, Gupta A, Ritchken P (2003) Hedging in the Possible Presence of Unspanned Stochastic Volatility: Evidence from Swaption Markets. J Finance 58:2219–2248
144. Fiorentina G, Sentana E, Shephard N (2004) Likelihood-Based Estimation of Latent Generalized ARCH Structures. Econometrica 72:1481–1517
145. Fisher M, Gilles C (1996) Estimating exponential-affine models of the term structure. Working Paper, Atlanta FED, Atlanta
146. Fleming MJ, Remolona EM (1999) Price Formation and Liquidity in the US Treasury Market: The Response to Public Information. J Finance 54:1901–1915
147. Forsberg L, Ghysels E (2007) Why Do Absolute Returns Predict Volatility So Well? J Finance Econom 5:31–67
148. French KR, Schwert GW, Stambaugh RF (1987) Expected stock returns and volatility. J Finance Econ 19:3–29
149. Fridman M, Harris L (1998) A Maximum Likelihood Approach for Non-Gaussian Stochastic Volatility Models. J Bus Econ Stat 16:284–291
150. Gallant AR, Hsieh DA, Tauchen GE (1997) Estimation of Stochastic Volatility Models with Diagnostics. J Econom 81:159–192
151. Gallant AR, Hsu C, Tauchen GE (1999) Using Daily Range Data to Calibrate Volatility Diffusions and Extract the Forward Integrated Variance. Rev Econ Stat 81:617–631
152. Gallant AR, Long JR (1997) Estimating stochastic differential equations efficiently by minimum chi-squared. Biometrika 84:125–141
153. Gallant AR, Rossi PE, Tauchen GE (1992) Stock Prices and Volume. Rev Finance Stud 5:199–242
154. Gallant AR, Tauchen GE (1996) Which Moments to Match. Econ Theory 12:657–681
155. Gallant AR, Tauchen G (1998) Reprojecting Partially Observed Systems With Application to Interest Rate Diffusions. J Ammer Stat Assoc 93:10–24
156. Garcia R, Lewis MA, Pastorello S, Renault E (2001) Estimation of Objective and Risk-neutral Distributions based on Moments of Integrated Volatility, Working Paper. Univer-

sité de Montréal, Banque Nationale du Canada, Università di Bologna, UNC

157. Garman MB, Klass MJ (1980) On the Estimation of Price Volatility From Historical Data. J Bus 53:67–78
158. Ghysels E, Harvey AC, Renault E (1996) Stochastic Volatility. In: Maddala GS, Rao CR (eds) Handbook of Statistics, vol 14. North Holland, Amsterdam
159. Ghysels E, Santa-Clara P, Valkanov R (2006) Predicting Volatility: How to Get the Most Out of Returns Data Sampled at Different Frequencies. J Econom 131:59–95
160. Glosten LR, Jagannathan R, Runkle D (1993) On the Relation Between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. J Finance 48:1779–1801
161. Gong FF, Remolona EM (1996) A three-factor econometric model of the US term structure. Working paper, Federal Reserve Bank of New York, New York
162. Gouriéroux C, Monfort A, Renault E (1993) Indirect Inference. J Appl Econom 8:S85–S118
163. Guo H, Neely CJ, Higbee J (2007) Foreign Exchange Volatility Is Priced in Equities. Finance Manag, forthcoming
164. Han B (2007) Stochastic Volatilities and Correlations of Bond Yields. J Finance 62:1491–1524
165. Hansen PR, Lunde A (2006) Realized Variance and Market Microstructure Noise. J Bus Econ Stat 24:127–161
166. Harvey AC (1998) Long memory in stochastic volatility. In: Knight J, Satchell S (eds) Forecasting Volatility in Financial Markets. Butterworth-Heinemann, London
167. Harvey AC, Ruiz E, Shephard N (1994) Multivariate Stochastic Variance Models. Rev Econ Stud 61:247–264
168. Harvey AC, Shephard N (1996) Estimation of an Asymmetric Stochastic Volatility Model for Asset Returns. J Bus Econ Stat 14:429–434
169. Heidari M, Wu L (2003) Are Interest Rate Derivatives Spanned by the Term Structure of Interest Rates? J Fixed Income 13:75–86
170. Heston SL (1993) A closed-form solution for options with stochastic volatility with applications to bond and currency options. Rev Finance Stud 6:327–343
171. Ho M, Perraudin W, Sørensen BE (1996) A Continuous Time Arbitrage Pricing Model with Stochastic Volatility and Jumps. J Bus Econ Stat 14:31–43
172. Huang X, Tauchen G (2005) The relative contribution of jumps to total price variation. J Finance Econom 3:456–499
173. Huang J, Wu L (2004) Specification Analysis of Option Pricing Models Based on Time-Changed Levy Processes. J Finance 59:1405–1440
174. Hull J, White A (1987) The Pricing of Options on Assets with Stochastic Volatilities. J Finance 42:281–300
175. Hurvich CM, Soulier P (2007) Stochastic Volatility Models with Long Memory. In: Andersen TG, Davis RA, Kreiss JP, Mikosch T (eds) Handbook of Financial Time Series. Springer,Berlin
176. Jung J (2006) Vexed by variance. Risk August
177. Jackwerth JC, Rubinstein M (1996) Recovering Probability Distributions from Option Prices. J Finance 51:1611–1631
178. Jacobs K, Karoui L (2007) Conditional Volatility in Affine Term Structure Models: Evidence from Treasury and Swap Markets. Working Paper, McGill University, Montreal
179. Jacobs K, Li X (2008) Modeling the Dynamics of Credit Spreads with Stochastic Volatility. Manag Sci 54:1176–1188
180. Jacquier E, Polson NG, Rossi PE (1994) Bayesian Analysis of Stochastic Volatility Models. J Bus Econ Stat 12:371–389
181. Jacquier E, Polson NG, Rossi PE (2004) Bayesian analysis of stochastic volatility models with fat-tails and correlated errors. J Econom 122:185–212
182. Jagannathan R, Kaplin A, Sun S (2003) An evaluation of multifactor CIR models using LIBOR, swap rates, and cap and swaption prices. J Econom 116:113–146
183. Jarrow R, Li H, Zhao F (2007) Interest Rate Caps "Smile" Too! But Can the LIBOR Market Models Capture the Smile? J Finance 62:345–382
184. Jiang GJ, Knight JL (2002) Efficient Estimation of the Continuous Time Stochastic Volatility Model via the Empirical Characteristic Function. J Bus Econ Stat 20:198–212
185. Jiang GJ, Tian YS (2005) The Model-Free Implied Volatility and Its Information Content. Rev Finance Stud 18:1305–1342
186. Johannes MS, Polson N (2003) MCMC Methods for Continuous-Time Financial Econometrics. In: Hansen LP, Aït-Sahalia I (eds) Handbook of Financial Econometrics. North-Holland, Amsterdam, forthcoming
187. Johannes MS, Polson N (2006) Particle Filtering. In: Andersen TG, Davis RA, Kreiss JP, Mikosch T (eds) Handbook of Financial Time Series. Springer, Berlin
188. Johnson H, Shanno D (1987) Option Pricing when the Variance Is Changing. J Finance Quant Anal 22:143–152
189. Jones CS (2003) Nonlinear Mean Reversion in the Short-Term Interest Rate. Rev Finance Stud 16:793–843
190. Jones CS (2003) The dynamics of stochastic volatility: evidence from underlying and options markets. J Econom 116:181–224
191. Joslin S (2006) Can Unspanned Stochastic Volatility Models Explain the Cross Section of Bond Volatilities? Working Paper, MIT, Cambridge
192. Joslin S (2007) Pricing and Hedging Volatility Risk in Fixed Income Markets. Working Paper, MIT, Cambridge
193. Kim SN, Shephard N, Chib S (1998) Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models. Rev Econ Stud 65:361–393
194. Lamoureux CG, Lastrapes WD (1994) Endogenous Trading vol and Momentum in Stock-Return Volatility. J Bus Econ Stat 14:253–260
195. Lee SS, Mykland PA (2006) Jumps in Financial Markets: A New Nonparametric Test and Jump Dynamics. Working Paper, Georgia Institute of Technology and University of Chicago, Chicago
196. Li H, Wells MT, Yu CL (2006) A Bayesian Analysis of Return Dynamics with Lévy Jumps. Rev Finance Stud, forthcoming
197. Li H, Zhao F (2006) Unspanned Stochastic Volatility: Evidence from Hedging Interest Rate Derivatives. J Finance 61:341–378
198. Liesenfeld R (1998) Dynamic Bivariate Mixture Models: Modeling the Behavior of Prices and Trading Volume. J Bus Econ Stat 16:101–109
199. Liesenfeld R (2001) A Generalized Bivariate Mixture Model for Stock Price Volatility and Trading Volume. J Econom 104:141–178
200. Liesenfeld R, Richard J-F (2003) Univariate and Multivariate Stochastic Volatility Models: Estimation and Diagnostics. J Empir Finance 10:505–531
201. Litterman R, Scheinkman JA (1991) Common Factors Affecting Bond Returns. J Fixed Income 1:54–61
202. Liu C, Maheu JM (2007) Forecasting Realized Volatility: A Bayesian Model Averaging Approach. Working Paper, University of Toronto, Toronto

203. Lo AW (1988) Maximum likelihood estimation of generalized Itô processes with discretely-sampled data. Econom Theory 4:231–247
204. Longstaff FA, Schwartz ES (1992) Interest Rate Volatility and the Term Structure: A Two-Factor General Equilibrium Model. J Finance 47:1259–1282
205. McAleer M, Medeiros MC (2007) Realized Volatility: A Review. Econom Rev, forthcoming
206. McFadden D (1989) A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration. Econometrica 57:995–1026
207. McQueen G, Vorkink K (2004) Whence GARCH? A Preference-Based Explanation for Conditional Volatility. Rev Finance Stud 17:915–949
208. Meddahi N (2001) An Eigenfunction Approach for Volatility Modeling. Working Paper, Imperial College, London
209. Meddahi N (2002) A Theoretical Comparison Between Integrated and Realized Volatility. J Appl Econom 17:475–508
210. Meddahi N (2002) Moments of Continuous Time Stochastic Volatility Models. Working Paper, Imperial College, London
211. Melino A, Turnbull SM (1990) Pricing foreign currency options with stochastic volatility. J Econom 45:239–265
212. Merton RC (1969) Lifetime portfolio selection under uncertainty: the continuous-time case. Rev Econ Stat 51:247–257
213. Merton RC (1973) An Intertemporal Capital Asset Pricing Model. Econometrica 41:867–887
214. Merton RC (1976) Option pricing when underlying stock returns are discontinuous. J Finance Econs 3:125–144
215. Merton RC (1980) On estimating the expected return on the market: An exploratory investigation. J Finance Econ 8:323–361
216. Mizrach B (2006) The Enron Bankruptcy: When Did The Options Market Lose Its Smirk. Rev Quant Finance Acc 27:365–382
217. Mizrach B (2007) Recovering Probabilistic Information From Options Prices and the Underlying. In: Lee C, Lee AC (eds) Handbook of Quantitative Finance. Springer, New York
218. Nelson DB (1991) Conditional Heteroskedasticity in Asset Returns: A New Approach. Econometrica 59:347–370
219. Nelson DB, Foster DP (1994) Asymptotic Filtering Theory for Univariate ARCH Models. Econometrica 62:1–41
220. Pan J (2002) The jump-risk premia implicit in options: evidence from an integrated time-series study. J Finance Econ 63:3–50
221. Parkinson M (1980) The Extreme ValueMethod for Estimating the Variance of the Rate of Return. J Bus 53:61–65
222. Pasquariello P, Vega C (2007) Informed and Strategic Order Flow in the Bond Markets. Rev Finance Stud 20:1975–2019
223. Pearson ND, Sun TS (1994) Exploiting the Conditional Density in Estimating the Term Structure: An Application to the Cox, Ingersoll, and Ross Model. J Finance 49:1279–1304
224. Pedersen AR (1995) A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. Scand J Stat 22:55–71
225. Pennacchi GG (1991) Identifying the Dynamics of Real Interest Rates and Inflation: Evidence Using Survey Data. Rev Finance Stud 4:53–86
226. Piazzesi M (2003) Affine Term Structure Models. In: Hansen LP, Aït-Sahalia I (eds) Handbook of Financial Econometrics. North-Holland, Amsterdam, forthcoming
227. Piazzesi M (2005) Bond Yields and the Federal Reserve. J Political Econ 113:311–344
228. Pitt MK, Shephard N (1999) Filtering via simulation: auxiliary particle filter. J Ammer Stat Assoc 94:590–599
229. Protter P (1992) Stochastic Integration and Differential Equations: A New Approach. Springer, New York
230. Renault E (1997) Econometric models of option pricing errors. In: Kreps D, Wallis K (eds) Advances in Economics and Econometrics, Seventh World Congress. Cambridge University Press, New York, pp 223–278
231. Richardson M, Smith T (1994) A Direct Test of the Mixture of Distributions Hypothesis: Measuring the Daily Flow of Information. J Financ Quant Anal 29:101–116
232. Rosenberg B (1972) The Behavior of Random Variables with Nonstationary Variance and the Distribution of Security Prices. working Paper, UCB, Berkley
233. Rubinstein M (1994) Implied Binomial Trees. J Financ 49:771–818
234. Santa-Clara P (1995) Simulated likelihood estimation of diffusions with an application to the short term interest rate. Dissertation, INSEAD
235. Schaumburg E (2005) Estimation of Markov processes with Levy type generators. Working Paper, KSM, Evanston
236. Schwert GW (1989) Why Does Stock Market Volatility Change Over Time? J Financ 44:1115–1153
237. Schwert GW (1990) Stock Volatility and the Crash of '87. Rev Financ Stud 3:77–102
238. Scott LO (1987) Option Pricing when the Variance Changes Randomly: Theory, Estimation and an Application. J Financ Quant Anal 22:419–438
239. Shephard N (1996) Statistical Aspects of ARCH and Stochastic Volatility Models. In: Cox DR, Hinkley DV, Barndorff-Nielsen OE (eds) Time Series Models in Econometrics, Finance and Other Fields. Chapman & Hall, London, pp 1–67
240. Shephard N (2004) Stochastic Volatility: Selected Readings. Oxford University Press, Oxford
241. Shiller RJ (1981) Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends? Ammer Econ Rev 71:421–436
242. Singleton KJ (2001) Estimation of affine asset pricing models using the empirical characteristic function. J Econom 102:111–141
243. Smith AA Jr (1993) Estimating Nonlinear Time-series Models using Simulated Vector Autoregressions. J Appl Econom 8:S63–S84
244. Stein JC (1989) Overreactions in the Options Market. J Finance 44:1011–1023
245. Stein EM, Stein JC (1991) Stock price distributions with stochastic volatility: an analytic approach. Rev Finance Stud 4:727–752
246. Tauchen GE (2005) Stochastic Volatility in General Equilibrium. Working Paper, Duke, University Durham
247. Tauchen GE, Pitts M (1983) The Price Variability-Volume Relationship on Speculative Markets. Econometrica 51:485–505
248. Tauchen GE, Zhou H (2007) Realized Jumps on Financial Markets and Predicting Credit Spreads. Working Paper, Duke University and Board of Governors, Washington D.C.
249. Taylor SJ (1986) Modeling Financial Time Series. Wiley, Chichester

250. Thompson S (2004) Identifying Term Structure Volatility from the LIBOR-Swap Curve. Working Paper, Harvard University, Boston
251. Todorov V (2006) Variance Risk Premium Dynamics. Working Paper, KSM, Evanston
252. Todorov V (2006) Estimation of Continuous-time Stochastic Volatility Models with Jumps using High-Frequency Data. Working Paper, KSM, Evanston
253. Trolle AB, Schwartz ES (2007) Unspanned stochastic volatility and the pricing of commodity derivatives. Working Paper, Copenhagen Business School and UCLA, Copenhagen
254. Trolle AB, Schwartz ES (2007) A general stochastic volatility model for the pricing of interest rate derivatives. Rev Finance Stud, forthcoming
255. Vasicek OA (1977) An equilibrium characterization of the term structure. J Finance Econ 5:177–188
256. Wiggins JB (1987) Option Values under Stochastic Volatility: Theory and Empirical Estimates. J Finance Econ 19:351–372
257. Whaley RE (1993) Derivatives on Market Volatility: Hedging Tools Long Overdue. J Deriv 1:71–84
258. Wright J, Zhou H (2007) Bond Risk Premia and Realized Jump Volatility. Working Paper, Board of Governors, Washington D.C.
259. Yang D, Zhang Q (2000) Drift-Independent Volatility Estimation Based on High, Low, Open, and Close Prices. J Bus 73:477–491
260. Zhang BY, Zhou H, Zhu H (2005) Explaining Credit Default Swap Spreads with the Equity Volatility and Jump Risks of Individual Firms. Working Paper, Fitch Ratings, Federal Reserve Board, BIS
261. Zhang L (2007) What you don't know cannot hurt you: On the detection of small jumps. Working Paper, UIC, Chicago
262. Zhang L, Mykland PA, Aït-Sahalia Y (2005) A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High Frequency Data. J Ammer Stat Assoc 100:1394–1411

Books and Reviews

Asai M, McAleer M, Yu J (2006) Multivariate Stochastic Volatility: A Review. Econom Rev 25:145–175

Bates DS (2003) Empirical option pricing: a retrospection. J Econom 116:387–404

Campbell JY, Lo AW, MacKinlay AC (1996) The Econometrics of Financial Markets. Princeton University Press, Princeton

Chib S, Omori Y, Asai M (2007) Multivariate Stochastic Volatility. Working Paper, Washington University

Duffie D (2001) Dynamic Asset Pricing Theory. Princeton University Press, Princeton

Gallant AR, Tauchen G (2002) Simulated Score Methods and Indirect Inference for Continuous-time Models. In: Hansen LP, Aït-Sahalia Y (eds) Handbook of Financial Econometrics. North-Holland, Amsterdam, forthcoming

Garcia R, Ghysels E, Renault E (2003) The Econometrics of Option Pricing. In: Hansen LP, Aït-Sahalia Y (eds) Handbook of Financial Econometrics. North-Holland, Amsterdam, forthcoming

Gouriéroux C, Jasiak J (2001) Financial Econometrics. Princeton University Press, Princeton

Johannes MS, Polson N (2006) Markov Chain Monte Carlo. In: Andersen TG, Davis RA, Kreiss JP, Mikosch T (eds) Handbook of Financial Time Series. Springer, Berlin

Jungbacker B, Koopman SJ (2007) Parameter Estimation and Practical Aspect of Modeling Stochastic Volatility. Working Paper, Vrije Universiteit, Amsterdam

Mykland PA (2003) Option Pricing Bounds and Statistical Uncertainty. In: Hansen LP, Aït-Sahalia Y (eds) Handbook of Financial Econometrics. North-Holland, Amsterdam, forthcoming

Renault E (2007) Moment-Based Estimation of Stochastic Volatility Models. In: Andersen TG, Davis RA, Kreiss JP, Mikosch T (eds) Handbook of Financial Time Series. Springer, Berlin

Singleton KJ (2006) Empirical Dynamic Asset Pricing: Model Specification and Econometric Assessment. Princeton University Press, Princeton

Structurally Dynamic Cellular Automata

ANDREW ILACHINSKI
Center for Naval Analyses, Alexandria, USA

Article Outline

Glossary

Adjacency matrix The *adjacency matrix* of a graph with N sites is an $N \times N$ matrix $[a_{ij}]$ with entries $a_{ij} = 1$ if i and j are linked, and $a_{ij} = 0$ otherwise. The adjacency matrix is symmetric ($a_{ij} = a_{ji}$) if the links in the graph are undirected.

Coupler link rules *Coupler rules* are local rules that act on pairs of *next-nearest sites* of a graph at time t to decide whether they should be linked at $t + 1$. The decision rules fall into one of three basic classes – totalistic (T), outer-totalistic (OT) or restricted-totalistic (RT) – but can be as varied as those for conventional cellular automata.

Decoupler link rules *Decoupler rules* are local rules that act on pairs of *linked sites* of a graph at time t to decide whether they should be unlinked at $t + 1$. As for coupler rules, the decision rules fall into one of three basic classes – totalistic (T), outer-totalistic (OT) or restricted-totalistic (RT) – but can be as varied as those for conventional cellular automata.

Degree The *degree* of a node (or site, *i*) of a graph is equal to the number of distinct nodes to which *i* is linked, and where the links are assumed to possess no directional information. In general graphs, the *in*-degree (= number of incoming links towards *i*) is distinguished from the *out*-degree (= number of outgoing links originating at *i*).

Effective dimension A quantity used to approximate the *dimensionality* of a graph. It is defined as the ratio between the average number of next-nearest neighbors to the average degree, both averaged over all nodes of the graph. The effective dimension equals the Euclidean dimension *d*, in cases where the graph is the familiar *d*-dimensional hypercubic lattice.

Graph A *graph* is a finite, nonempty set of nodes (referred to as "sites" throughout this article), together with (a possibly empty) set of edges (or links). The links may be either *directed* (in which case the edge from a site *i*, say, is directed away from *i* toward another site *j*, and is considered distinct from another directed edge originating at *j* and pointed toward *i*) or *undirected* (in which case if a link exists between sites *i* and *j* it carries no directional information).

Graph grammar *Graph grammars* (sometimes also referred to as *graph rewriting systems*) apply formal language theory to networks. Each language specifies the space of "valid structures", and the production (or "rewrite") rules by which given graphs may be transformed into other valid graphs.

Graph metric function The *graph metric function* defines the distance between any two nodes, *i* and *j*. It is equal to the length of the shortest path between *i* and *j*. If no path exists (such as when *i* and *j* are on two disconnected components of the same graph), the distance is assumed to be equal to ∞.

Graph-rewriting automata *Graph-rewriting automata* are generalized CA-like systems in which both (the number of) nodes and links are allowed to change.

Next-nearest neighbor Two sites *i* and *j* are *next-nearest neighbors* in a graph if (1) they are not directly linked (so that $a_{ij} = 0$; see *adjacency matrix*), and (2) there exists at least one other site *k* such that $k \notin \{i, j\}$, and *i* and *j* are both lined to *k*.

Random dynamics approximation The long-term behavior of structurally dynamic cellular automata may be approximated in certain cases (in which the structure and value configurations are both sufficiently random and uncorrelated) by a *random dynamics approximation*: values of sites are replaced by the probability p_σ of a site having value σ (and is assumed to be equal for all sites), and links between sites are replaced by the probability p_ℓ of being linked (and also assumed to be the same for all pairs of sites). The approximation often yields qualitatively correct predictions about how the real system evolves under a specific set of rules; for example, to predict whether one expects unbounded growth or that the lattice will eventually settle onto a low periodic state or simply decay.

Restricted totalistic rules *Restricted totalistic rules* are a generalized class of link rules (operating on pairs of sites, *i* and *j*), analogous to "outer totalistic" rules (that operate on site values) used in conventional CA. The local neighborhood around *i* and *j* is first partitioned into three sets: (1) the two sites, *i* and *j*; (2) sites connected to either *i* or *j*, but not both; and (3) sites connected to both *i* and *j*. The *restricted totalistic rule* is then completely defined by associating a specific action with each possible 3-tuple of site-value sums (where the individual components represent a unique sum in each of the three neighborhoods).

Structurally dynamic cellular automata *Structurally dynamic cellular automata* are generalizations of conventional cellular automata models in which the underlying lattice structure is dynamically coupled to the local site-value configurations.

SDCA model hierarchy The *SDCA model hierarchy* is a set of eight related structurally dynamic cellular automata models, defined explicitly for studying their formal computational capabilities. The hierarchy is ordered (from lowest to highest level) according to their relative computational strength. For example, the SDCA model at the top of the hierarchy is capable of simulating a conventional CA with a speedup factor of two.

Definition of the Subject

Structurally dynamic cellular automata (abbreviated, SDCA) are a generalized class of CA in which the topological structure of the (usually quiescent) underlying lattice is dynamically coupled to the local site value configuration. The coupling is defined to treat *geometry* and *value configurations* on an approximately equal footing: the lattice structure is altered locally as a function of individual site neighborhood value-states and geometries, while the underlying local topology supports site-value evolution precisely as in conventional nearest-neighbor CA models defined on random lattices.

SDCA provide a dynamical framework for a CA-like analysis of the generation, transmission and interaction of topological disturbances in a lattice. Moreover, they provide a natural testbed for studying self organized geometry; by which we mean true structural evolution, and not merely space-time patterns of value configurations that may be *interpreted* geometrically (but are really just "bits" of information overlayed on top of an otherwise static background lattice).

Introduction

SDCA were formally introduced in 1986 as part of a physics doctoral dissertation by Ilachinski [31], and developed further by Ilachinski and Halpern [29,30], Halpern [21,23], Halpern and Caltagirone [22], Majercik [39], and Alonso-Sanz and Martín [6,7,8]; in their original incarnation [28], and at least two subsequent papers [22,61], SDCA were called *topological automata*. Pedagogical discussions appear in Adamatzky [1] and Ilachinski [32]. Extensions of the basic SDCA model (all discussed in this article) include the addition of probabilistic rules, memory and reversibility.

Applications include the simulation of crystal growth [36], the study of pattern formation of random cellular structures [66], modeling synaptic plasticity in neural network models [19], phase transitions in chemical systems [62], chemical self-assembly [26], and gene-regulatory networks [22]. Majercik [39] has studied SDCA as generalized models of computation, and describes a CA-universal SDCA that can simulate any conventional CA of the same dimension.

More recently, O'Sullivan [56] and Saidani [63,64] have used graph-based CA models similar to SDCA to study urban dynamics and emergent behaviors of self-reconfigurable robots, respectively. Tomita et al. [67,68,69,70,71] have introduced *graph-rewriting automata* in which both links and (the number of) nodes are allowed to change; and show that these systems are capable of both self-replication and Turing universality (among with many other emergent behaviors). Since SDCA provide the basic formalism for describing locally induced topological changes within arbitrary graphs, they are a potentially powerful general tool for studying complex adaptive networks, such as communication and social networks [6]. The *concept* behind SDCA has also been used as a foundation for philosophical musings about computationally emergent artificiality [49].

More ambitious applications of SDCA encroach on fundamental physics. Because SDCA are inherently self-modifying systems – in which physical events are not just dynamically coupled to, but are an integral part of the spatio-temporal arena on which their transformations are defined – they are a potentially powerful methodological and ontological tool for exploring discrete pre-geometric theories of space-time [42]. Just as "value structure" solitons are ubiquitous in conventional CA models [32,74], "link structure" solitons might emerge in SDCA; physical particles would, in such a scheme, be viewed as geometrodynamic disturbances propagating within a dynamic lattice. Three SDCA-like theories of pregeometry have recently been proposed in which space-time is a self-organized emergent construct: Hillman [27], Nowotny and Requardt [55] and Wolfram [75].

Finally, we briefly comment on ostensible overlaps between SDCA and four other related fields of study: (1) *Lindenmeyer* (or L-) *systems*, (2) *graph grammars*, (3) *random graphs* (abbreviated, RG), and (4) *dynamic network analysis* (abbreviated, DNA). L-systems [57] are generalized CA systems in which the number of sites can grow with time, and consist of recursive rules for rewriting strings of symbols. If interpreted graphically, abstract symbol strings can be used to model growth processes of plants and evolving morphology of physical organisms. Graph grammars [20,35] apply formal language theory to networks, and consist of production rules that define the set of "valid structures" in a given graph language. The study of RG [15] was introduced by Erdos and Renyi in the late 1950s [16], and is a mathematical framework for exploring the general topological structures of computational systems and the behavior of certain random dynamical systems. Like SDCA, RG describes evolving graphs, but the dynamics are global and random. DNA [41,50] is an emerging field that fuses traditional social network theory with statistical analysis and modeling; part of its charter is to explore general properties of network generation and evolution.

While, conceptually speaking, there is a prima facie relationship between SDCA and all four fields of study, the elucidation of a more precise nature of the relationship between SDCA and these other systems awaits a future study. (The relationship appears particularly strong between SDCA and a generalized L-system called the *graph development system* (abbreviated, GDS), introduced by Doi [14], but not developed further since its original conception. Using incidence matrices to represent arbitrary topologies, GDS is essentially a grammar by which submatrices of the whole matrix are rewritten to describe topological changes. SDCA also formally falls under the broader rubrics of DNA and RG; however, there is no explicit reference to SDCA in the current literature of either field.)

The Basic Model

Conventional CA are defined on fixed, and typically regular, lattices (one-dimensional lines, two-dimensional Euclidean or hexagonal grids, etc.), the sites of which are populated with discrete-valued dynamic elements ($\equiv \sigma_i \in \{0, 1, \dots, k\}$, where i labels a particular site on the lattice) that evolve according to local transition functions, $f: \sigma_i \rightarrow \sigma_i'$. We emphasize that the dynamics of conventional CA are confined to the temporal evolution of the σ_is.

SDCA generalize conventional CA in two ways: (1) they relax the assumption that the underlying lattice is uniform, allowing the local site ↔ site connectivity pattern to vary throughout the lattice; and (2) they allow *both* the set $\{\sigma_i\}$ *and* the lattice to evolve according to local transition rules. The most obvious – also the most dramatic – conceptual change this entails over the dynamics of conventional CA, is that the meaning of "local" itself changes as a function of how the SDCA system evolves: previously far separated sites may become neighbors; and sites that are local at time t may become far separated at some later time, t'.

To properly define SDCA, we first generalize regular lattices to mathematical graphs $\equiv G$ () possessing arbitrary topology. Assuming G has N lattice sites, and that G is (for now) an *undirected* graph (meaning that none of G's links carry directional information), G is completely defined by the N-by-N *adjacency* matrix, ℓ_{ij}:

$$\ell_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are linked;} \\ 0 & \text{otherwise.} \end{cases} \qquad (1)$$

Using the graph metric function,

$$D_{ij} = \underset{\text{Paths}, P_{ij}}{\text{Minimum}} \left[\text{\#links}, l_{rs} \,|\, \{r, s\} \in P_{ij} \right], \qquad (2)$$

we can write a general r-neighborhood CA *value*-transition rule 'f' (which will from now on refer generically to as a σ-rule) in the form

$$\sigma_i^{t+1} = f\left[\{\sigma_j^t\} | j \in S_r^G(i) \right], \qquad (3)$$

where $S_r^G(i) = \{j | D_{ij} \leq r\}$ is the radius-r *graph sphere* about the site i. In words, the value of σ_i^{t+1} is some function, f, of the values σ_j^t in radius r graph sphere around the site i. With this distance measure, G becomes a *discrete metric space*. If G is a one-dimensional line, and $r = 1$, then $S_r^G(i) = \{i-1, i, i+1\}$; i. e., it is equal to the conventional three site local neighborhood of elementary CA.

We now formally extend a conventional CA's dynamic arena – limited to the values $\sigma_i^t \in \{0, 1, \dots, k-1\}$, $i = 1, \dots, N$ – to one that includes the components of the underlying lattice's adjacency matrix:

$$\begin{cases} \sigma^{t+1} = F_\sigma[\{\sigma^t\}, \{\ell^t\}] \\ \ell^{t+1} = F_\ell[\{\sigma^t\}, \{\ell^t\}] \end{cases}, \qquad (4)$$

where F_σ and F_ℓ are some functions (to be defined explicitly below) that explicitly couple the changing *value* states and *geometries*. The complete system at time t is specified by the state-vector

$$|G\rangle_t = |\sigma_1^t, \dots, \sigma_N^t;\ \{\ell_{ij}^t\}\rangle\,. \qquad (5)$$

The time-evolution of $|G\rangle$ proceeds according to the following transition rules: (i) σ-rules of the general form given above and familiar from *CA* simulations and (ii) ℓ-rules, which are divided into site *couplers*, linking previously unconnected vertices and site *decouplers*, which disconnect linked points. Because the topology can be altered only by either a deletion of *existing links* or an addition of links between pairs of vertices 'i' and 'j' with $D_{ij} = 2$, the dynamics is *strictly local*.

To be more precise, we first restrict the general σ-rule F_1 to (maximally symmetric) *totalistic* (**T**) and *outer-totalistic* (**OT**) type. Since the underlying lattice is a fully dynamic object, $|G\rangle$ will, in general, tend towards having a complex local geometry with an unspecified local *directionality*. The most general rules which can therefore be applied are those which are completely invariant under all rotation and reflection symmetry transformations on local neighborhoods. **T** (**OT**) σ-rules are then specified by listing particular *sums* $\{\alpha\}$ (*outer-sums* $\{\alpha_0\}, \{\alpha_1\}$ corresponding to center site values '0' and '1' respectively) for which the value of the center site becomes '1'. Formally,

$$\sigma_i^{t+1} = \phi_{\{\alpha\}} \left(\sum_j \ell_{ij}^t \sigma_j^t, \sigma_i^t \right), \qquad (6)$$

where

$$\phi_{\{\alpha\}}(x, a) = \begin{cases} \sum_\alpha \delta(x + a, \alpha) & \longleftrightarrow \mathbf{T} \\ a \sum_{\alpha_1} \delta(x, \alpha_1) + (1 - a) \sum_{\alpha_0} \delta(x, \alpha_0) & \longleftrightarrow \mathbf{OT}\,, \end{cases} \qquad (7)$$

and $\delta(x, y)$ is the *Kronecker* delta. Note that $\sum_j \ell_{ij}^t \sigma_j^t$ sums the values of all sites 'j' linked to 'i' at time 't'. The action on the state $|G\rangle$ is represented by

$$\widehat{\phi}^i_{\{\alpha\}} |\sigma\rangle_t = \left| \sigma_1^t, \dots, \sigma_i^{t+1} = \phi_{\{\alpha\}} \left(\sum \ell_{ij}^t \sigma_j^t, \sigma_i^t \right), \dots, \sigma_N^t \right\rangle, \qquad (8)$$

where we distinguish the *operator* $\widehat{\phi}^i$ acting on the global value state from the actual local transition *function* ϕ which transforms each site value.

Link Rules

Local geometry altering rules are constructed by direct analogy: for any two selected sites i and j we restrict attention to site values of vertices contained within a 1-*sphere* of either site; that is, to all $k \in S_1(i,j) = S_1(i) \cup S_1(j)$. Link operators, whose action on the state is represented by:

$$\begin{aligned} \textit{decouplers:}\quad & \widehat{\psi}^{ij}_{\{\beta\}} \left|\ell^t_{ij}\right\rangle = \left|\ell^t_{11}, \ldots, \ell^{t+1}_{ij} = \psi^{ij}, \ldots, \ell^t_{NN}\right\rangle \\ \textit{couplers:}\quad & \widehat{\omega}^{ij}_{\{\varepsilon\}} \left|\ell^t_{ij}\right\rangle = \left|\ell^t_{11}, \ldots, \ell^{t+1}_{ij} = \omega^{ij}, \ldots, \ell^t_{NN}\right\rangle , \end{aligned} \tag{9}$$

either *link* or *unlink* two sites '*i*' and '*j*' depending on whether the actual sum of values in $S_1(i,j)$ matches any of those given in the $\{\beta\}$ or $\{\varepsilon\}$ lists, which completely define *decouplers* and *couplers*, respectively.

In order to construct classes of rules analogous to the two types of σ-rules defined above, we partition the local neighborhood into 3 disjoint sets (see Fig. 1): $S_1(i,j) = V_{ij} \cup A_{ij} \cup B_{ij}$, where

$$\begin{cases} \vee_{ij} = \{i,j\}, \\ A_{ij} = \{k | k \in C_1(i) \cap C_1(j)\}, \text{ where } C_1(i) = S_1(i) - \{i\}, \\ B_{ij} = S_1(i) \cup S_1(j) - \vee_{ij} - A_{ij} . \end{cases} \tag{10}$$

The action of link operators is then conveniently expressed as a function of the sums within the individual partitions. Defining $\nu_{ij} = \sigma_i + \sigma_j$, $a_{ij} = \sum_{k \in A_{ij}} \sigma_k$, and $b_{ij} = \sum_{k \in B_{ij}} \sigma_k$, we get *decouplers*, $\psi^{ij}_{\{\beta\}} = \psi^{ij}_{\{\beta\}}(\nu_{ij}, a_{ij}, b_{ij})$, where

$$\psi^{ij}_{\{\beta\}}(x,y,z) = \begin{cases} \{1 - \sum_k \delta(x+y+z, \beta_k)\}\ell_{ij} & \leftrightarrow \mathbf{T} \\ \{1 - \sum_k \delta(x, \beta_{1,k})\delta(y+z, \beta_{2,k})\}\ell_{ij} & \leftrightarrow \mathbf{OT} \\ \{1 - \sum_k \delta(x, \beta_{1,k})\delta(y, \beta_{2,k})\delta(z, \beta_{3,k})\}\ell_{ij} & \leftrightarrow \mathbf{RT} , \end{cases} \tag{11}$$

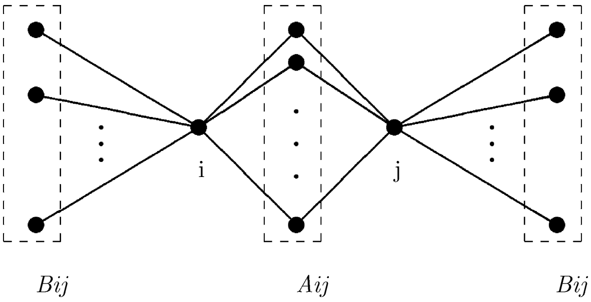
and *couplers*, $\omega^{ij}_{\{\varepsilon\}} = \omega^{ij}_{\{\varepsilon\}}(\nu_{ij}, a_{ij}, b_{ij})$, where

$$\omega^{ij}_{\{\varepsilon\}}(x,y,z) = \begin{cases} \delta(D_{ij}, 2) \sum_k \delta(x+y+z, \varepsilon_k) & \leftrightarrow \mathbf{T} \\ \delta(D_{ij}, 2) \sum_k \delta(x, \varepsilon_{1,k})\delta(y+z, \varepsilon_{2,k}) & \leftrightarrow \mathbf{OT} \\ \delta(D_{ij}, 2) \sum_k \delta(x, \varepsilon_{1,k})\delta(y, \varepsilon_{2,k})\delta(z, \varepsilon_{3,k}) & \leftrightarrow \mathbf{RT} . \end{cases} \tag{12}$$

In the above expressions, **RT** stands for *restricted totalistic* rules which maximally subdivide the local neighborhood. The inclusion of an ℓ_{ij} in the expressions for ψ assures that only those sites already *linked* can be decoupled and the $\delta(D_{ij}, 2)$ in the equations defining ω are put in to make sure that only sites separated by distance = 2 may be dynamically coupled.

The three type-specific *sums* appearing above are indexed with the following conventions:

- **T** rules are defined by the '*k*' overall sums of values in $S_1(i,j)$ for which the particular action is to be taken. For example, define 'ψ' by *unlinking* '*i*' and '*j*' if the total sum = 1 (= β_1), 3 (= β_2) or 5 (= β_3). Equation (11) then states that $\ell^{n+1}_{ij} = 0$ if and only if $\ell^n_{ij} = 1$ and $\nu^n_{ij} + a^n_{ij} + b^n_{ij} \in \{1,3,5\}$.
- **OT** rules are specified by giving '*k*' 2-tuples $(\beta_{1,k}, \beta_{2,k})$, and $(\varepsilon_{1,k}, \varepsilon_{2,k})$, where $\{1,k\}$ labels the sum '$\sigma_i + \sigma_j$' and $\{2,k\}$ labels the corresponding *outer sum* $= \sum_{s \in S_1(i,j) - \{i,j\}} \sigma_s$. For example, *link* '*i*' and '*j*' if $\sigma_i + \sigma_j = 0$ and outer sum = $\{3,4\}$, so that 'ω' is defined by listing the two 2-tuples $(\varepsilon_{1,1}0, \varepsilon_{2,1} = 3)$ and $(\varepsilon_{1,2} = 0, \varepsilon_{2,2} = 4)$.
- **RT** rules are completely specified by giving the '*k*' 3-tuples of values $(x\sigma_i + \sigma_j, y = \text{sum in } A, z = \text{sum in } B)$, for which the link operation between '*i*' and '*j*' is to be performed. For example, define 'ψ' by *unlinking* '*i*' and '*j*' for the following values of partitioned sums: (0, 0, 1), (0, 0, 2), (0, 1, 1), (1, 1, 1); we then have that $(\beta_{1,1} = 0, \beta_{2,1} = 0, \beta_{3,1} = 1)$, $(\beta_{1,2} = 0, \beta_{2,2} =$



Structurally Dynamic Cellular Automata, Figure 1
Neighborhood partitioning. In the same way as *outer* sites can be considered separately for σ-transitions, we may, for topology transitions, distinguish between those sites belonging to *both* i and j (∈ A_{ij}) and those belonging to *one* of the two sites but not both (∈ B_{ij}). In this way we obtain the analogous *totalistic* (T), *outer-totalistic* (OT), and an additional type called *restricted totalistic* (RT)

$0, \beta_{3,2} = 2)$, $(\beta_{1,3} = 0, \beta_{2,3} = 1, \beta_{3,3} = 1)$, and $(\beta_{1,4} = 0, \beta_{2,4} = 1, \beta_{3,4} = 1)$.

Global transition operators are obtained by applying individual σ- and ℓ- operators to all *sites* and *site-pairs* in the graph G:

$$\begin{cases} \widehat{\Phi}_{\{\alpha\}}\,|\sigma\rangle = \prod_i \widehat{\phi}^{\,i}_{\{\alpha\}}\,|\sigma\rangle \ , \\ \widehat{\Psi}_{\{\beta\}}\,|\ell\rangle = \prod_{n\langle ij\rangle} \widehat{\psi}^{\,ij}_{\{\beta\}}\,|\ell\rangle \ , \\ \widehat{\Omega}_{\{\varepsilon\}}\,|\ell\rangle = \prod_{nn\langle ij\rangle} \widehat{\omega}^{\,ij}_{\{\varepsilon\}}\,|\ell\rangle \ , \end{cases} \tag{13}$$

where the products for $\widehat{\Psi}$ and $\widehat{\Omega}$ need to be taken only over *nearest* and *next nearest* pairs respectively. Given the full *value-topology* transition rule Γ, defined by

$$|G\rangle_{t+1} = (\widehat{\Omega}\widehat{\Psi}\widehat{\Phi})|G\rangle_t = \Gamma|G\rangle_t \ , \tag{14}$$

the fundamental problem is to understand the generic behavior of accessible graphs-G emerging from all possible initial structures and value configurations. We emphasize that the lattice *fully* participates in the dynamics and that, in general, no embedding is implied – it is the abstract *connectivity* itself whose evolution we are attempting to trace.

An Example

The application of the rather cumbersome expressions defining transition rules is in practice extremely straightforward, as we demonstrate with the following example: Consider a graph G defined as a (5×5) lattice with some distribution of values $\sigma = 1$ at time '$t = 1$' (see Fig. 2). We are interested in one *global* update of the system $|G\rangle_{t=1} \xrightarrow{\Gamma} |G\rangle_{t=2}$ with rules specified by

$$\begin{array}{l} \text{(value)} \\[1em] \text{(topology)} \\ \end{array} \left\{ \begin{array}{l} \Phi_{\{\alpha\}} : \{\alpha\}_{\mathbf{T}} = \{\alpha_1 = 1, \alpha_2 = 3, \alpha_3 = 5\} \ , \\ \Psi_{\{\beta\}} : \{\beta\}_{\mathbf{OT}} = \left\{ \begin{array}{l} (\beta_{1,1} = 1, \beta_{2,1} = 3) \\ (\beta_{1,2} = 1, \beta_{2,2} = 4) \end{array} \right\} , \\ \Omega_{\{\varepsilon\}} : \{\varepsilon\}_{\mathbf{OT}} = \{\varepsilon_{1,1} = 1, \varepsilon_{2,1} = 3\} \ . \end{array} \right. \tag{15}$$

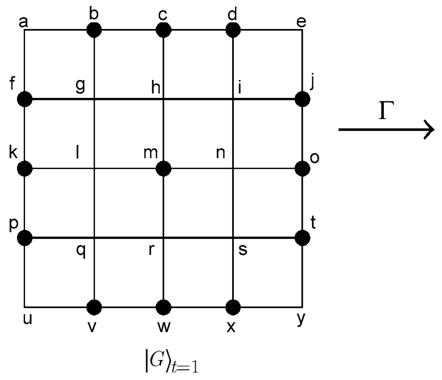
We evolve the system by systematically sweeping through all *sites*, *linked* pairs, and *next-nearest neighbors*:

1. *All Sites:* … setting $\sigma_i = 1$ only at those 'i' for which the sum of the values at 'i' and its neighbors is equal to '2' at $t = 1$. By "neighbors" of any point 'i' we will always mean the set of vertices linked to 'i': (a, b), (h, m) and (x, y), for example, are all neighbors at $t = 1$. Writing out a few value-changing terms explicitly, we find that

$$\begin{aligned} \sigma_c^{t=2} &= \phi\left(\sigma_b^{t=1} + \sigma_c^{t=1} + \sigma_d^{t=1} + \sigma_h^{t=1}\right) \\ &= \phi(3) = 1, \quad \text{and} \\ \sigma_b^{t=2} &= \phi\left(\sigma_a^{t=1} + \sigma_b^{t=1} + \sigma_c^{t=1} + \sigma_g^{t=1}\right) \\ &= \phi(2) = 0. \end{aligned} \tag{16}$$

2. *All linked pairs of sites 'i' and 'j':* … removing those links only if the 2-tuple $(\alpha, \beta) \in \{(1, 3), (1, 4)$, where $\alpha = \sigma_i + \sigma_j$ and 'β' is the sum of values of the neighbors of 'i' and 'j' at $t = 1$. For the points 'c' and 'h', for example, we have $(\alpha, \beta) = (1, 3)$, so that the link ℓ_{ch} is no longer present in $|G\rangle_{t=2}$:

$$\begin{aligned} \ell_{ch}^{t=2} = \psi\Big(&\sigma_c^{t=1} + \sigma_h^{t=1}, \sigma_b^{t=1} + \sigma_d^{t=1} + \sigma_g^{t=1} \\ &+ \sigma_i^{t=1} + \sigma_m^{t=1}\Big)\, \ell_{ch}^{t=1} \\ = \psi&(1, 3)(1) = 0 \,. \end{aligned} \tag{17}$$



Structurally Dynamic Cellular Automata, Figure 2

Sample dynamic update of a (5 × 5) lattice from $t = 1$ to $t = 2$, obeying a T-type σ-rule with $\sigma \to \sigma'$ for local sums = 1, 3, 5 (i. e. $\alpha \in \{1, 3, 5\}$), and OT-type ℓ-rules: (i) *link* for $\{\varepsilon_{1,1} = 1, \varepsilon_{2,1} = 3\}$ and (ii) *unlink* for $\{\beta_{1,1} = 1, \beta_{2,1} = 3\}$ and $\{\beta_{1,2} = 1, \beta_{2,2} = 4\}$. *Solid* sites indicate that $\sigma = 1$

3. *All next-nearest neighbors 'i' and 'j': …* linking them only if the 2-tuple $(\alpha, \beta) = \{(1, 3)\}$. By "next-nearest neighbor" we mean those pairs which are themselves unlinked but which share at least one other linked neighbor: (a, g), (h, r) and (w, y), for example, are all next-nearest neighbors at $t = 1$. For 'c' and 'g' we find

$$\begin{aligned}\ell_{cg}^{t=2} &= \omega\Big(\sigma_c^{t=1} + \sigma_g^{t=1}, \sigma_b^{t=1} + \sigma_d^{t=1} + \sigma_f^{t=1} \\ &\qquad + \sigma_h^{t=1} + \sigma_l^{t=1}\Big)\,\delta(D_{cg}, 2) \\ &= \omega(1, 3)(1) = 1\,. \end{aligned} \tag{18}$$

Notice that although $\ell_{dn}^{t=1} = 0 \to \ell_{dn}^{t=2} = 1$, it is hidden by overlap with the remaining links $\ell_{di}^{t=2} = 1$ and $\ell_{in}^{t=2} = 1$. For this reason, not all link changes can always be observed directly in the following figures.

Other sites and links are updated in precisely the same manner. Had the link-rules been of **T**-type, only one sum would have to be considered: the sum of the values of the points in question along with their neighbors' values. Had they been, instead, of **RT**-type, three sums would have to be considered: the sum of the values of the sites in question, the sum of the values of their common neighbors (neighborhood A in Fig. 1) and the sum of the values of the points that are neighbors of one of the considered points, but not of the other (neighborhood B in Fig. 1). The final state $|G\rangle_{t=2}$ emerges after the above process has been applied concurrently to all pairs, neighbors and next-nearest neighbors in $|G\rangle_{t=1}$.

Comments

We conclude this section by making a few important general comments:

Comment 1. As defined above, Γ consists of three operators acting simultaneously on the state $|G\rangle$. More generally, one may prescribe any of 10 possible time-orderings to the operators Ω, Ψ and Φ. That is, specify certain intermediate state dependencies, so that, for example $\Gamma_1|G\rangle \equiv (\Omega\Psi)(\Phi|G\rangle)$ would in general be expected to yield results different from, say, $\Gamma_2|G\rangle \equiv \Omega(\Phi(\Psi|G\rangle))$. While we will be solely concerned with the synchronous time ordering defined above, we do not expect the qualitative results to depend critically on this choice.

Comment 2. A given rule Γ is completely defined by the set of sums $\{\alpha\}$, $\{\beta\}$ and $\{\varepsilon\}$. Alternatively, we can conveniently summarize a chosen transition rule by its vector-code $\vec{C} = (c[\phi], c[\psi], c[\omega])_{a,b}$, where

$$\begin{aligned} c[\phi] &= \begin{cases} \sum_\alpha 2^\alpha & \leftrightarrow \mathbf{T} \\ \sum_{\alpha_0} 2^{2\alpha_0} + \sum_{\alpha_1} 2^{(2\alpha_1+1)} & \leftrightarrow \mathbf{OT} \end{cases} \\ c[\psi] &= \begin{cases} \sum_k 2^{\beta_k} & \leftrightarrow \mathbf{T} \\ \sum_k 2^{3\beta_{2,k}+\beta_{1,k}} & \leftrightarrow \mathbf{OT} \\ \sum_k 2^{3(\beta_{2,k}+a\beta_{3,k})+\beta_{1,k}} & \leftrightarrow \mathbf{RT} \end{cases} \\ c[\omega] &= \begin{cases} \sum_k 2^{\varepsilon_k} & \leftrightarrow \mathbf{T} \\ \sum_k 2^{3\varepsilon_{2,k}+\varepsilon_{1,k}} & \leftrightarrow \mathbf{OT} \\ \sum_k 2^{3(\varepsilon_{2,k}+b\varepsilon_{3,k})+\varepsilon_{1,k}} & \leftrightarrow \mathbf{RT} \end{cases} \end{aligned} \tag{19}$$

where $a = \max\{\beta_{2,k}\} + 1$, $b = \max\{\varepsilon_{2,k}\} + 1$, and must be specified only for **RT**-type topology rules. The Γ appearing in the above example, therefore, can be summarized by $c[\phi] = 42$, $c[\psi] = 2^{3(3)+1} = 1024$ and $c[\omega] = 2^{3(4)+1} + 2^{3(3)+1} = 9216$. Note that '$\Psi$' and '$\Omega$' are chosen always to be of the same type.

Comment 3. Computer simulations of these systems require that some measures be taken to prevent possible memory overflows, such as would happen in cases either of pure coupling, where links are continually added and none deleted, or in isolated regions of a graph where for a few sites more neighbors are added than are allowed by memory. We thus introduce *working* link transition rules

$$\tilde{\psi}^{ij} \equiv \begin{cases} \psi^{ij} & \longleftrightarrow d_i' \text{ or } d_j' > \delta \equiv d_{\min} \\ 1 & \longleftrightarrow \text{else}\,, \end{cases} \tag{20}$$

$$\tilde{\omega}^{ij} \equiv \begin{cases} \omega^{ij} & \longleftrightarrow d_i' \text{ or } d_j' < \Delta \equiv d_{\max} \\ 0 & \longleftrightarrow \text{else}\,, \end{cases} \tag{21}$$

where $d_i =$ degree(i) (i. e. number of neighbors of i). In words: make a sweep of the lattice, temporarily storing the candidates to *add* and *delete* for each point. If, for any point i, the updated degree is greater than δ then proceed with deleting the stored deletion-candidates, otherwise do not delete; similarly, provided that the updated degree is less than Δ proceed with addition. Thus, it is sufficient that *one* of two points allow a dynamic link change between them for that change to be enacted. In the following, the complete constrained dynamics will be quoted as $\vec{C}_{(a,b)}^{[\delta,\Delta]}$. If constraints play no role in the actual evolution of specific examples, they will be left out of the definition.

Comment 4. Because each dynamic update involves three separate types of processing, the number of possible rules is extraordinarily large (see Table 1). Unlike pure σ-transitions, however, the fraction of the total number which yield interesting behavior (i. e. neither immediately explosive, where the number of links increases with-

Structurally Dynamic Cellular Automata, Table 1

Numbers of possible rules for each of the three types of transition rules. $d =$ *maximum allowable degree* and $a =$ *maximum sum to be used from partition* A_{ij}*. Example:* for $d = 5$, we have $N_\phi = 4096$, $N_\psi = 2^{24} \sim 2 \times 10^7$ and $N_\omega = 2^{21} \sim 2 \times 10^6$. We thus have $N_T = N_\phi N_\psi N_\omega \sim 10^{17}$ possible type OT Γ s

Rule type	ϕ	ψ	ω
T	2^{d+1}	2^{2d}	2^{2d-1}
OT	2^{2d+2}	$2^{6(d-1)}$	$2^{3(2d-3)}$
RT	—	$2^{3(a+1)(2d-1)}$	$2^{3(a+1)(2d+1)}$

out bound, nor immediately degenerative, where an initial graph rapidly dwindles to a few isolated links) appears to be manageably smaller.

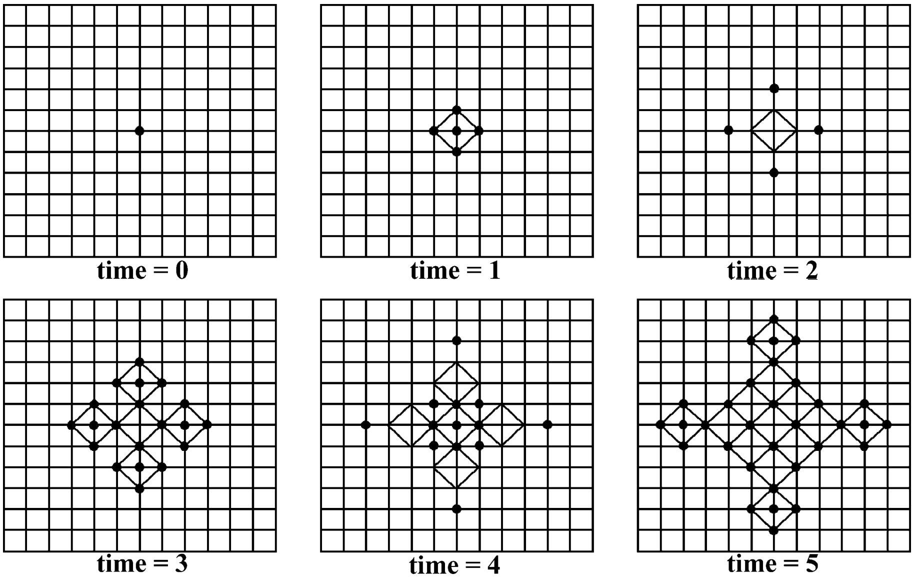
Comment 5. Although it is the intrinsic geometrical patterning whose generic behavioral properties we are trying to deduce, one may approach SDCA from an alternative point of view: maintain the emphasis on unraveling the value configurational behavior, and interpret the presence of $[\Psi, \Omega]$ as background operators inducing nonlocal spatial connectivities. Whereas the systems defined above are completely abstract entities, in that locality is strictly defined by the link structure, the alternative scheme would be to embed the discrete networks in some specified manifold, and to study the effects of dynamically allocated nonlocal communication channels.

Emerging Patterns and Behaviors

Consider patterns that emerge from simple value seeds starting from ordered two dimensional Euclidean lattices. A single non-zero site may represent a small local disturbance that then propagates outward, restructuring the lattice. With appropriately chosen Γs one can induce a rich spectrum of different time evolutions only slightly perturbed by very few concurrent link changes to ones in which the initial geometry becomes radically altered. (The graphical representation of evolving one dimensional systems, in which link additions must be shown as *arcs* to avoid overlap with existing links, is needlessly confusing and is not considered.)

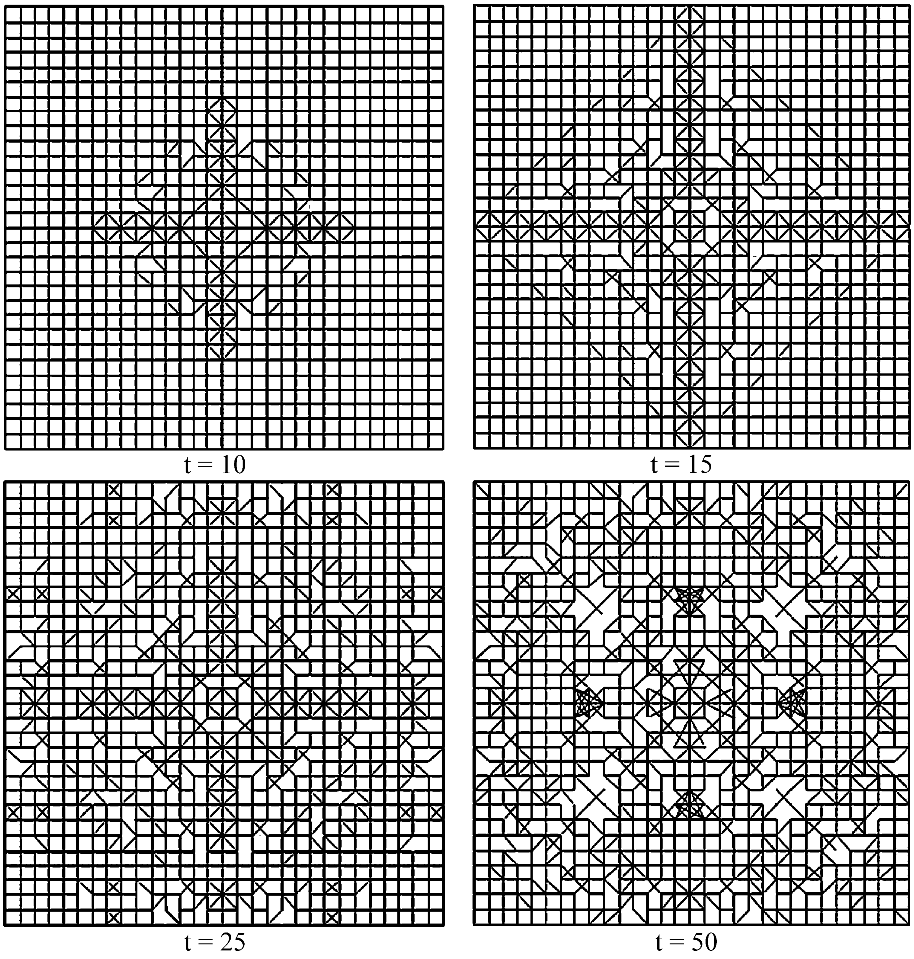
Figure 3 shows the first five iterations of a system starting from a four neighbor lattice with a single non-zero site at its center, the link structure is given explicitly and the solid circles represent sites with $\sigma = 1$. Notice how the link additions *follow* the emerging corrugated boundary surface of the value configuration. Remember that link additions are more than passive markers indicating particular correlations between local value configurations and structure; their presence directly influences all subsequent value development in their immediate vicinity.

Figure 4 (in which site values are suppressed for clarity) shows the continued development of this system. Though boundary effects begin to appear by $t = 25$, the



Structurally Dynamic Cellular Automata, Figure 3

First five iterations of an SDCA system starting from a 4-neighbor Euclidean lattice seeded with a single non-zero site at the center. The global transition rule Γ consists of T σ-rule and RT ℓ-rules: $\vec{C} = (26, 69648, 32904)_{[3,3]}$ (see text for rule definitions and code). Solid sites have $\sigma = 1$



Structurally Dynamic Cellular Automata, Figure 4
Several further time frames in the structural evolution of the same system shown in the preceding figure. The values have been suppressed for clarity. The boundaries of the original lattice do not extend beyond the region shown so that the development is strictly confined to a 31 × 31 graph

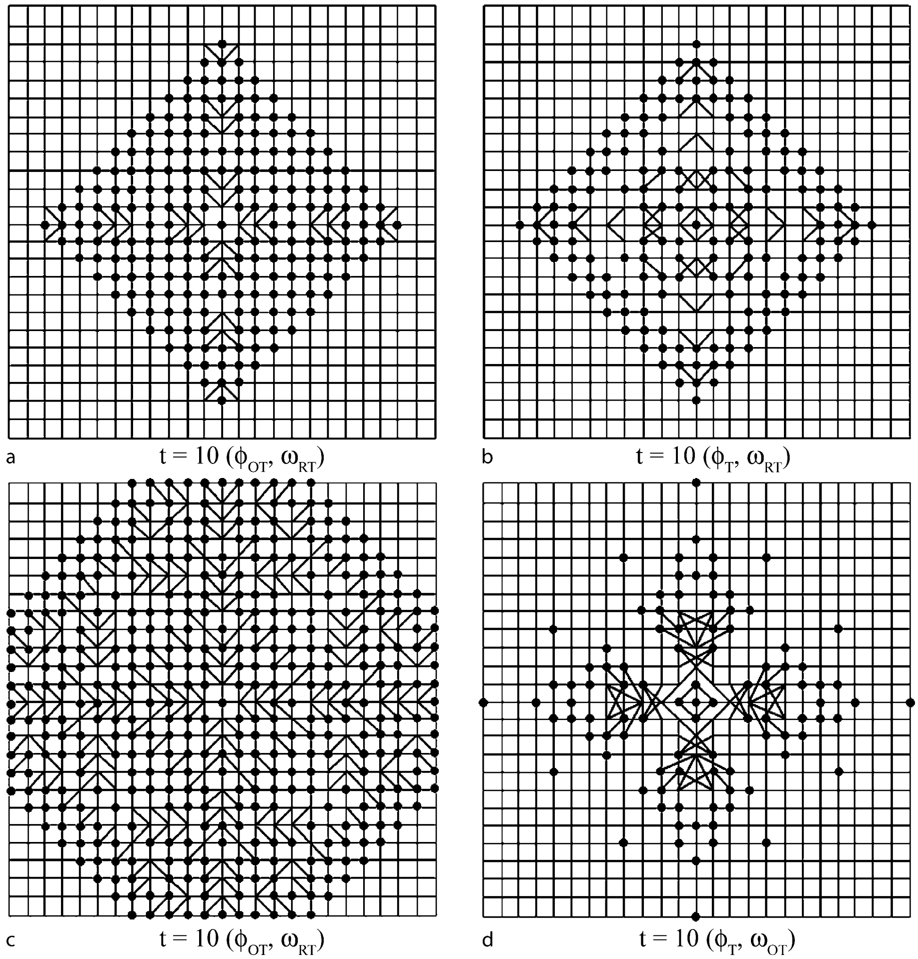
characteristic manner in which this particular Γ restructures the initial graph is clear:

- There is a high degree of geometrical organization (the symmetry of the initial state is trivially preserved by the totally symmetric Γ).
- The lattice remains connected.
- The distribution of link changes made throughout the lattice remains fairly uniform (i. e. there is an approximate uniformity in the probability of appearance of particular local value states which induce a structural change.
- Link-lengths do not get arbitrarily large.

The last point implies that for a system embedded in the plane, communication channels remain approximately local. The global pattern emerges as a consequence of local ordering. On the other hand, Γs for which link-lengths get arbitrarily large are also easy to find.

Some other varieties of behavior are shown in Figs. 5 and 6. Figures 5a and b are representative of the class of ℓ-rules that only mildly perturb the underlying lattice (and for which σ states do not differ much from their conventional CA cousins). Other rules, of course, may have a stronger effect on the lattice, giving rise to associated σ states bearing little or no resemblance to their conventional CA counterparts.

Figure 5c shows an example of a link rule that *accelerates* the outward propagation of the value configuration. Compare the diameter of this pattern to that in the earlier figures, both shown at equal times. The outwardly oriented


a
t = 10 (ϕ_{OT}, ω_{RT})
b
t = 10 (ϕ_{T}, ω_{RT})
c
t = 10 (ϕ_{OT}, ω_{RT})
d
t = 10 (ϕ_{T}, ω_{OT})

Structurally Dynamic Cellular Automata, Figure 5

Snapshot views of four typical developing states starting from a single non-zero site at the center of a 4-neighbor graph. Γs are as follows: a OT $c[\phi] = 1022$ and RT coupler $c[\omega] = 16$, $b = 1$; b T $c[\phi] = 22$ and RT coupler $c[\omega] = 32$, $b = 2$; c OT $c[\phi] = 1022$ and RT coupler $c[\omega] = 8$, $b = 1$; d T σ- and OT ℓ-rules $\vec{C} = (682, 19634061312, 133120)_{[2,8]}$

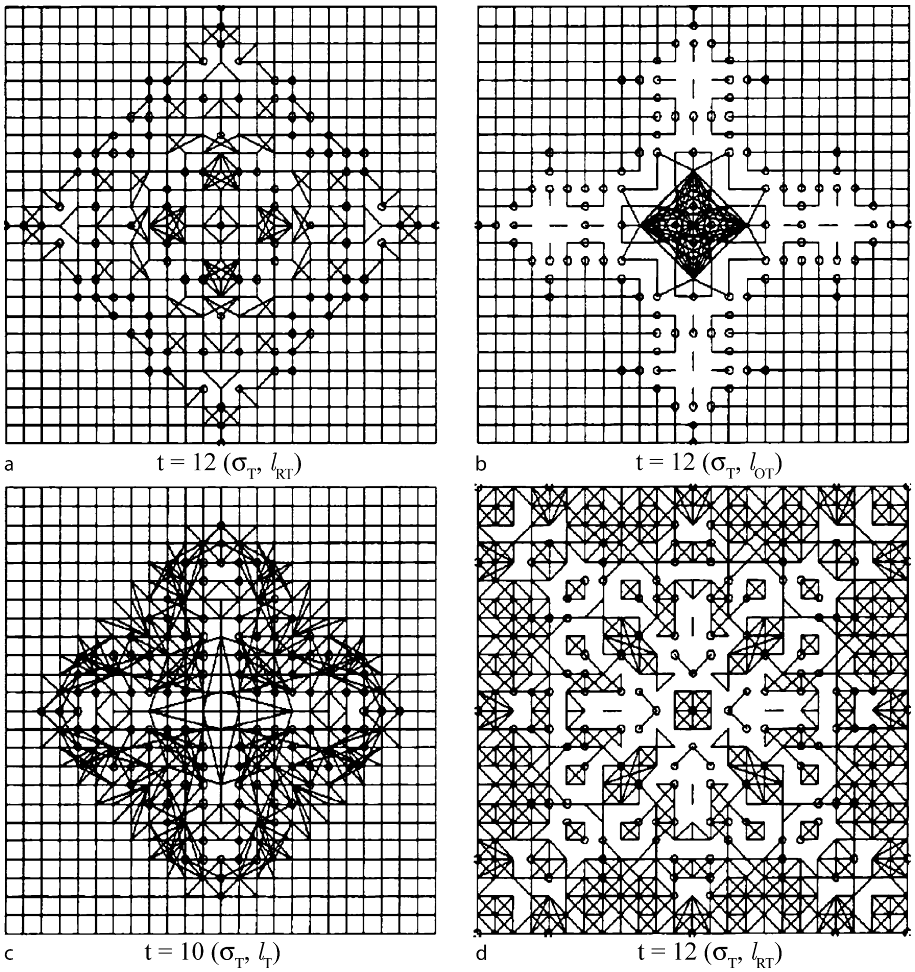
links that emerge from sites along the boundary surface become conduits by which non-zero values rapidly propagate. Had the underlying lattice topology been suppressed in this figure, and attention focused exclusively on the developing σ state, we could have interpreted the result as showing an effective increase in information propagation speed due to non-local connectivities (see comment 5 of the previous section).

Figure 5d, on the other hand, gives an example in which the link dynamics *lags behind* the σ development. The boundary proceeds outward essentially unaffected by changes in geometry, which are themselves confined to the interior parts of the lattice (at least at this early stage of this system's development).

Figure 6 shows snapshot views of a few system undergoing a slightly more complex evolution. Figure 6b, for example, shows a rule in which the outward σ propagation rapidly deletes most links from the original lattice but leaves a complex (though structurally stable) geometry at the origin of the initial disturbance. Figure 6c, on the other hand, shows a typical state of a system whose global connectivity becomes progressively more complicated.

A typical evolution starting from an initial state in which all sites are randomly assigned $\sigma = 1$ with probability $p = 1/2$ is shown in Fig. 7. Notice the rapid development of complex local connectivity patterns, the appearance of which points to a geometrical self-organization.

In general, structural behaviors emerging from random σ-states under typical Γs can be grouped into four basic classes (not to be confused with Wolfram's classification of elementary CA [74]):


a
$t = 12\ (\sigma_T, l_{RT})$
b
$t = 12\ (\sigma_T, l_{OT})$
c
$t = 10\ (\sigma_T, l_T)$
d
$t = 12\ (\sigma_T, l_{RT})$

Structurally Dynamic Cellular Automata, Figure 6

Four more examples of states emerging from simple seeds. Figure a, b, c start from 4-neighbor graphs and d from an 8-neighbor graph (≡ 4-neighbor with diagonals). Γs are as follows: a T σ- and RT ℓ-rules $\vec{C} = (42, 69648, 32904)_{[3,3]}$; b T σ- and OT ℓ-rules $\vec{C} = (42, 589952, 8192)_{[2,8]}$; c T σ- and ℓ-rules $\vec{C} = (42, 128, 4)_{[0,10]}$; d T $c[\phi] = 682$ and RT ℓ-rules defined explicitly by $\Psi_{(104),(114),(124),(103),(113),(123)}$ and $\Omega_{(111),(215)}$

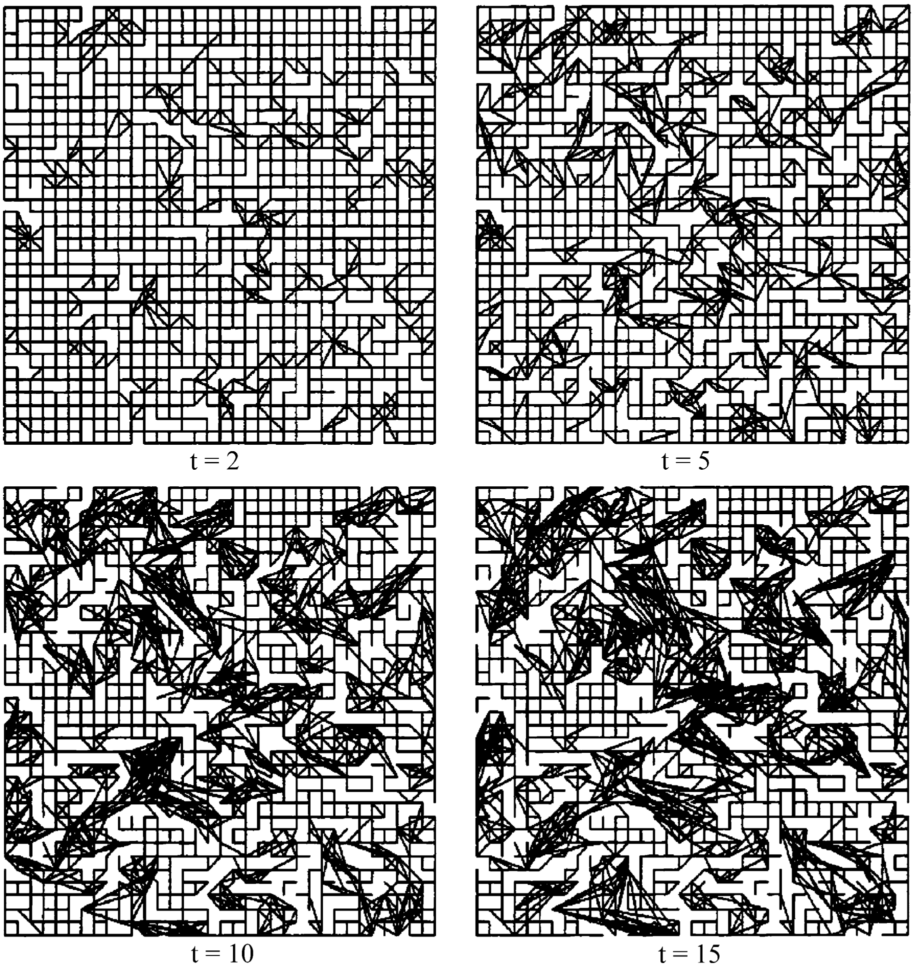
- *Class-1*, in which initial graphs decay into structurally much simpler final states: most links are destroyed, and graphs ℓ^t_{ij}, for sufficiently large t, consist essentially of a large number of small local subgraphs.
- *Class-2*, whose final states are characterized by periodic but globally connected geometries. SDCA typically arise in this class either because of a specific class-2 Φs remaining unchanged by the coupling to the lattice or class-3 Φs coupling with $\{\Psi, \Omega\}$ in such a way as to induce a lattice structure that supports a periodic state.
- *Class-3*, consisting of SDCA that tend to grow in size and complexity, at least as measured by two basic metrics: the *average degree*, $\langle\text{deg}\rangle \equiv (1/N) \cdot \sum_i \left[|S_1(i)| - 1\right]$, and *effective dimensionality*, $D_{\text{effec}} \equiv \langle N_{\text{nn}}\rangle / \langle\text{deg}\rangle$, where $\langle N_{\text{nn}}\rangle$ is the average number of next-nearest neighbors. The values of both $\langle\text{deg}\rangle$ and D_{effec} increase without bound for class-3 SDCA (unless an arbitrary upper constraint Δ is imposed on Γ).

Because the σ-density responds to the changing local neighborhood structure, it is possible that what at first appears to be an explosive growth in fact eventually leads to a more sedate, if not static, behavior at some larger $\langle\text{deg}\rangle \gg \Delta$. Φs that yield $\langle\sigma\rangle_t \sim$ *constant* over a range of $\langle\text{deg}\rangle$ (such as the *sum modulo*-2 rule; see below), when coupled with link rules that themselves become progressively less active with increasing $\langle\text{deg}\rangle$, may induce evolutions leading to only mild changes within specific ranges of the local structural parameters.

- *Class-4*, which is a provisional class (pending stronger evidence) that denotes a set of rules that yield open-ended σ- and ℓ changes, but during which the value of D_{effec} remains roughly constant. Ψs and Ωs belonging to this class effectively induce a *structural equilibrium*: despite the fact that large numbers of link changes continue to be made, so that the detailed structure of the evolving graph continually changes, the average ratio of the number of next-nearest to nearest neighbors stays approximately constant over long periods of time. While there is evidence to suggest this class is *real*, simulations have unfortunately been run for too short a time and on graphs containing too few sites to permit making any conclusive statements regarding the veracity of this class. Nonetheless, it is tempting to speculate that, for arbitrary values of D^*, there exists at least one set of SDCA rules for which $D_{\text{effec}} \approx D^*$ (within a desired $\epsilon > 0$) as the size of the graph $N \to \infty$. (*Pseudo* class-4 behavior, of course, can always be artificially induced either by imposing severe $[\delta, \Delta = \delta]$ constraints, or, as must typically be done for category-3 Γs, by deliberately impeding growth with some threshold Δ.)

Statistical Measures

As evidenced by Fig. 7, it is already nontrivial to meaningfully visualize the short-time evolution of (initially) *regu-*



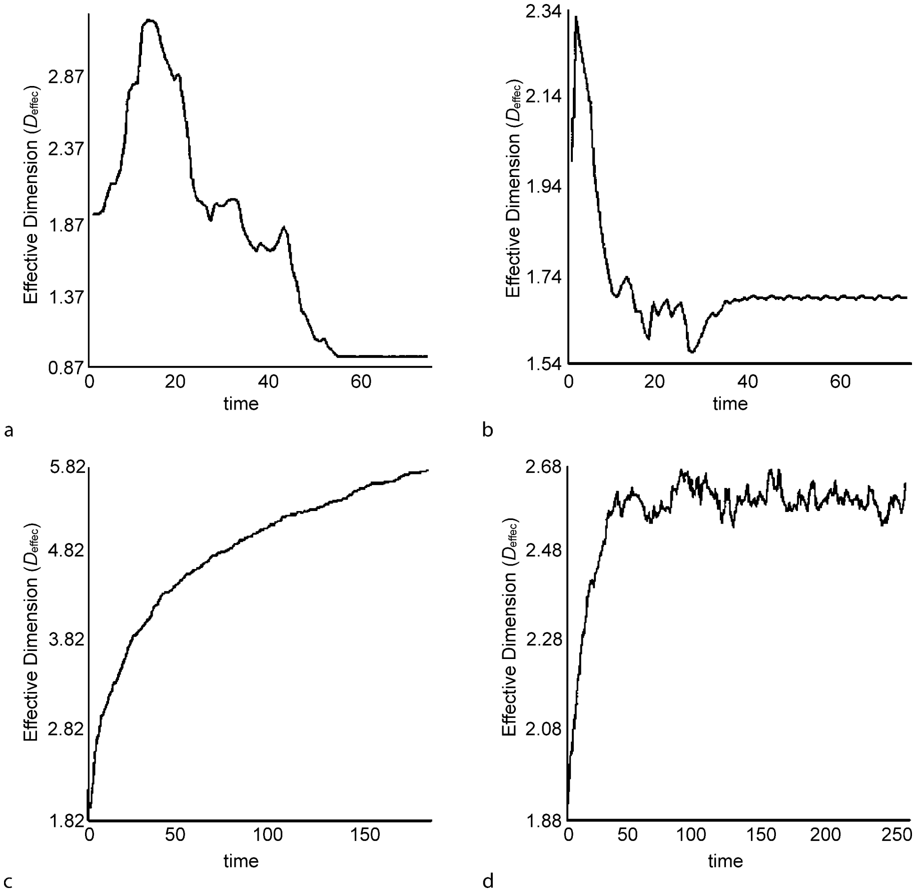
Structurally Dynamic Cellular Automata, Figure 7
Evolution of a 35 × 35 lattice, with randomly seeded sites. The development proceeds according to T σ- and OT ℓ-rules defined by code $\vec{C} = (84, 36864, 2048)$. The constraints are $[\delta = 0, \Delta = 10]$. The appearance of localized substructures is evidence of a geometrical self-organization

lar lattices that start with random initial value state. Visualizing the long-term dynamics of systems that start from a completely random state is even more difficult (although graph visualization algorithms may help). However, even in cases for which a direct visual inspection of the dynamics reveals little, one can always indirectly keep abreast of a given system's properties by monitoring its core structural and behavioral measures (a more detailed account is given in [31]).

Site value measures include the average density of sites with value $\sigma = 1$, $\langle \sigma \rangle_t \sim (1/N) \sum_{i=1}^{N} \sigma_i^t$; the local value correlation, $C^t \equiv \langle \sigma_i^t \cdot \sigma_j^t \rangle - (\rho^t)^2$, where $\langle \sigma_i^t \cdot \sigma_j^t \rangle$ is averaged over all pairs i and j with $\ell_{ij} = 1$; the fraction of sites whose value *changes* during one step of the evolution, $\Delta_t \equiv (1/N) \sum_{i=1}^{N} \{\sigma_i^{t-1} \oplus_2 \sigma_i^t\}$, where $\oplus_2$ is a sum *modulo*-2.

Geometry measures include the *average degree*, $\langle \text{deg} \rangle$; the *average number of next-nearest neighbors*, $\langle N_{\text{nn}} \rangle_t \equiv (1/N) \sum_i [|S_2(i)| - |S_1(i)| - 1]$; and D_{effec}. A measure of how the actual size of local neighborhoods changes with time may be obtained by embedding graphs into the two-dimensional plane and calculating the average pathlength at time t. Of course, global features that describe all complex networks – such as *connectivity*, *density*, *clustering*, and *path lengths* (2,6), are applicable to SDCA as well.

Link changes may be monitored by keeping track of (1) the total number of *link changes* (allowed under prescribed constraint conditions), $\Delta_t^{(l)} \equiv (1/2) \sum_{i=1}^{N} \cdot \sum_{j=1}^{N} \{l_{ij}^t \oplus_2 l_{ij}^{t-1}\}$; (2) the *constraint influence*, $f_l \equiv \Delta_t^{(l)} / N_t^{(l)}$, where $N_t^{(l)}$ is the total number of link changes that would have occurred in the absence of constraints ($f_l = 1$ indicates that the evolution is *pure*,



Structurally Dynamic Cellular Automata, Figure 8

Time development of the effective dimensionality D_{effec} for each of the four categories of behavior (see text): a T type Γ defined by $\vec{C} = (42, 128, 4)$; b T σ- and OT ℓ-rules $\vec{C} = (64, 9216, 1024)$; c T σ- and OT ℓ-rules $\vec{C} = (682, 512, 512)_{[0,10]}$; d T σ- and RT ℓ-rules defined explicitly by $\vec{\beta} \in \{(011), (110), (121), (233), (243)\}$ and $\vec{\epsilon} \in \{(120), (010), (021), (224)\}$

meaning it is unaffected by constraints; $f_l \sim$ *small* suggests that the imposed constraint window $[\delta, \Delta]$ has resulted in observed structures that are *impure*); (3) the link *creation-* and link *deletion-ratios*, $f_C \equiv N_C/\Delta_t^{(l)}$ and $f_D \equiv N_D/\Delta_t^{(l)}$, where N_C and N_D are the numbers of link created and destroyed, respectively; (4) the *activity levels*, $\gamma_C^t \equiv N_C/N_{\text{nn}}^{t-1}$ and $\gamma_D^t \equiv N_D/N_{\text{l}}^{t-1}$ (where N_l^{t-1} is the number of links at time $t-1$), which give the number of dynamic alterations relative to the corresponding spaces from which the candidates for alteration are selected; and (5) the *link evolution index*, $\gamma_L^n \equiv \left(1/N_l^{t=0}\right) \sum_i \sum_j \{l_{ij}^{t=0} \oplus_2 l_{ij}^n\}$, which gives the fraction of the initial lattice remaining after n iterations.

Figure 8 shows time series plots of D_{effec} for rules in each of the four behavioral classes defined above. The initial structure in each case is 35 × 35 4-neighbor Euclidean lattice, so that $D_{\text{effec}}^{t=0} \sim 2$. Figure 8a gives an example of class-1 behavior, in which a short period of initial growth is followed by a decay into mostly disconnected clusters. The final state is characterized by $\langle \text{deg} \rangle < 1$, and is stable. Figure 8b shows a system that starts from the same initial state as in Fig. 8a but whose Γ leads to a periodic geometry. Just the right number of links have been deleted to permit regions with isolated activity to emerge.

Figure 8c shows class-3 behavior in which D_{effec} steadily increases. The apparent leveling off seen toward the end of the run is due both to a decreased overall activity level and the increasingly effect of the $\Delta = 10$ constraint. The system in Fig. 8d exhibits class-4 behavior, characterized by a ongoing structural development within a relatively narrow interval of values of D_{effec}. Note that the structural changes here are essentially *pure*, and are not merely artifacts of any imposed constraints. Ilachinski (3) explores a wide range of emergent behaviors across all four classes, and examines the qualitative relationship between emergent behavior and initial σ- and ℓ-seeding.

Phase Plots

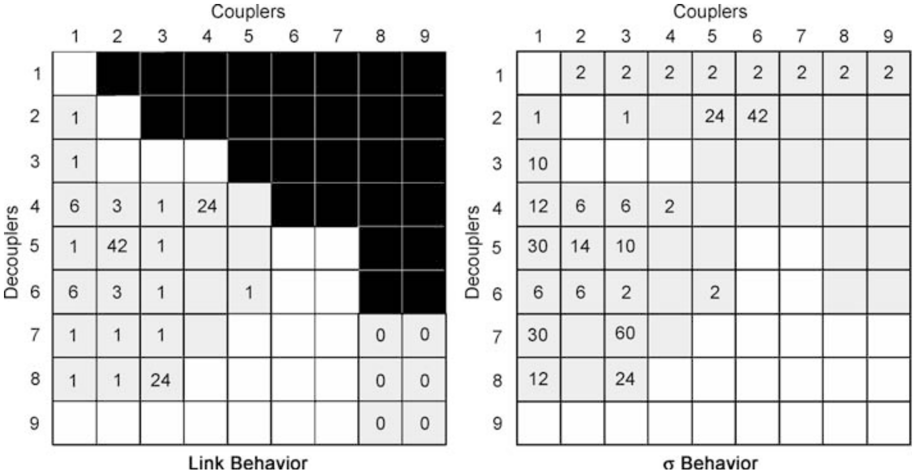
While it is of obvious interest to systematically explore every possible combination of β's and ϵ's that define ℓ-rules, Table 1 unfortunately suggests that the resulting rule space is simply too large. Nonetheless, we can learn much even by focusing our attention on a small subset of the complete rule space, keeping Φ, the initial σ-seeding, and all other factors constant. Specifically, consider the subset of all possible ℓ-rules that consists of **OT** link rules consisting of a single *coupler*, ω, and a single *decoupler*, ψ. Moreover, let $\phi \equiv \oplus_2$ (i. e., sum *modulo*-2 rule), demand that only pairs of $\sigma = 0$ sites be considered for a link change, and consider ℓ-rules belonging to the following set:

$$\begin{array}{lll} \textit{decouplers:} & \{\beta_{1,1} = 0, \beta_{2,1} = \mu\}, & 1 \leq \mu \leq 0, \\ \textit{couplers:} & \{\epsilon_{1,1} = 0, \epsilon_{2,1} = \nu\}, & 1 \leq \nu \leq 0. \end{array} \tag{22}$$

Figure 9 summarizes the behavior of a four neighbor, 25 × 25 lattice with periodic boundary conditions, starting from an initial σ-seed consisting of a single nonzero site. Four basic kinds of structural behaviors emerge:

1. *Static state:* this trivially occurs when the link rules are unable to take effect; namely, when $\mu \geq 7$ and $\nu \geq 8$.
2. *Rapid growth:* for an entire range of μ and ν, the average number of neighbors for each site of the lattice increases rapidly for 20–30 iterations.
   This number would likely continue to increase, were it not for the constraint conditions ($\equiv [0, 10]$). The "final state" is neither stable nor periodic. One sometimes also sees *delayed growth* in this class of behavior, in which case the link structure is initially relatively quiescent (and the behavior of the system as a whole mimics that of a conventional CA). As coupler rules are triggered by specific σ states, the average degree of the lattice rapidly increases (at least until the constraint conditions take effect).
3. *Spontaneous decay:* when decouplers are stronger than couplers, the average degree typically decreases. If this occurs too rapidly, the structure surrounding the single nonzero valued site may become isolated from other parts of the lattice. If a few non-zero values do not leak out into the outlying regions, link changes remain confined to the central subgraph, leading to either rapid stability or periodicity.
4. *Initial growth, followed by periodicity:* this is the least common behavior, and requires a delicate balance between coupler and decoupler rules.

It is interesting to compare these results with those obtained from a random σ seed. In this case, the sharp divisions between characteristic behaviors disappear, and there is a pronounced increase in the number of links for all μ and ν. However, the inclusion of an additional decoupler, may induce decay and periodicity. For example, consider the same initial lattice and Φ as used in Fig. 9, fix two **OT** ℓ-rules $\Psi_{(0,5)}$ and $\Omega_{(0,1)}$, and add the decoupler $\Psi_{(0,\mu)}$: $\{\beta_{1,1} = 0, \beta_{2,1} = \mu\}$, $1 \leq \mu \leq 9$. Surveying the emergent behaviors for this range of μ's, one now finds decaying lattices for $\mu \geq 2$. In each case, the initial graph succumbs to periodicity following a transient of between 50 and 100 iterations. The evolving lattice is also more prone to break up into small disconnected subgraphs.



Structurally Dynamic Cellular Automata, Figure 9

Phase plot **that summarizes behavior of a four neighbor, 25 × 25 lattice with periodic boundary conditions, starting from an initial σ-seed consisting of a single nonzero site. Γ is defined by the sum *modulo*-2 σ-rule and ℓ-rules of the form: *decouplers*–$\{\beta_{1,1} = 0, \beta_{2,1} = \mu\}$, *couplers*–$\{\epsilon_{1,1} = 0, \epsilon_{2,1} = \nu\}$. *Grey areas* in both plots denote *periodic states*. *White areas* denote *growth* in the plot for link behavior, and a *nonperiodic* state for σ-behavior. The *black* area that appears in the link-bevavior plot denotes *decay*. Numbers that appear in *individual boxes* denote period lengths**

Although, just as in conventional CA, small changes to ℓ-rules can lead to large differences in emergent behavior, they generally appear to do so in a more predictable and patterned manner. Of course, particular classes of Γ may induce more complex phase plots; for example, isolated pockets of anomalous (and rapidly shifting) behavior may appear within larger surrounding regions undergoing otherwise mutually consistent and slowly changing dynamics. A better sense of the space of possible emergent behaviors, along with a deeper understanding of the relationship between Φ and ℓ-rules, awaits a future study.

SDCA as Models of Computation

The basic SDCA model, as outlined above (which we will denote as SDCA$_0$ to avoid possible confusion with the hierarchy of related SDCA models introduced in this section), was modified and generalized by Majercik [39] into a form more suitable for addressing its formal *computational capabilities* rather than as an exploratory toolkit for describing physical processes (which is the primary reason for which SDCA$_0$ were first conceived). Motivated primarily by finding models of human brain function (for which one intuitively expects nonlocal neural connections to play a fundamental role in the rewiring of neural tissue), Majercik shows that suitably generalized SDCA are not only capable of universal computation, but actually represent a more efficient class of computational models than conventional CA. Majercik also reports an SDCA that can solve the *firing squad* problem in $O(\log t)$ time (i. e., exponentially faster than the $O(t)$ in conventional CA), and a class of CA-*universal* SDCA models that can simulate any conventional CA with a speedup factor of two. (The *firing squad* problem [46] consists of finding a rule for which all sites in a CA evolve into a special state after the exactly the same number of steps.)

Majercik proceeds by first identifying five properties of SDCA$_0$ that, while reasonable from a physical modeling standpoint, make it difficult to rigorously formulate and prove theorems:

1. *Finiteness:* The requirement that SDCA$_0$ be strictly finite, both in time and space, is obviously necessary for computer experiments, but is unnecessarily restrictive for general theorem proving. Likewise, the assumption that the sets α, β and ϵ must be finite is questioned.
2. *Bidirectionality:* While SDCA$_0$ are defined with symmetric links, an obvious generalization that makes the basic model more readily applicable to neural dynamics (among other kinds of physical and biological systems) is to allow for unidirectional links.
3. *Link-rule Asymmetry:* While SDCA$_0$'s link decoupler function (Eq. (11)) contains the factor ℓ_{ij} to explicitly prevent the system from inadvertently linking two unlinked sites, SDCA$_0$ does not include an analogous term for the coupler function (that is, a term to prevent an evolving system from inadvertently unlinking two linked sites).

4. *Inconsistency:* While σ-rules effectively ignore site positions, all three types of link rules assume that the various neighborhoods surroundings individual sites (A_{ij}, B_{ij}, and $C_{ij} \equiv \{k | D_{ik} = 1 \oplus D_{jk} = 1\}$, where '$\oplus$' denotes *exclusive or*) are all *recognized as such* by the dynamics. That is, the link rules effectively " know" the positions of a site's neighbors, while σ-rules possess no such information.
5. *Small Rule Set:* The class of σ- and ℓ-rules used by SDCA_0 may be generalized to include a far broader class of transition functions.

On the basis of these observations, Majercik [39] introduces a set of three core models to define a hierarchy of eight alternative SDCA computational systems, $\{\text{SDCA}^{(1)}, \text{SDCA}^{(2)}, \ldots, \text{SDCA}^{(8)}\}$. The three core models are (1) the *relative location* model ($= M_{\text{R}}$), (2) the *labeled links* model ($= M_{\text{L}}$), and (3) the *symmetric links* model ($= M_{\text{S}}$). They differ only in the degree to which their σ- and ℓ-transition functions depend on *specific* sites. For example, M_{R}'s transition functions depend on the state and exact relative position of each neighbor (and therefore "knows" the exact source of any state in a local neighborhood). In M_{L}, links are labeled and the transition functions know both neighbor states and the label of the links to given neighbors, but the exact neighbor locations remain unspecified. Finally, in M_{S}, it is assumed that no information about the source of the neighborhood states exists, and transition functions only know the number of neighbors in a particular state.

Each of the three core models may be defined in two versions: an *unbounded links* (abbreviated, UL) version, in which the number of neighbors a given site can have is unbounded, and a *bounded links* (abbreviated, BL) version, in which an explicit upper limit is imposed. In addition, there is also one *finite labels* version of M_{L}. Majercik imposes certain mild conditions on the local transition functions; for example, that local neighborhoods always remain strictly finite, σ-rules leave quiescent neighborhoods alone, and that links between sites with quiescent neighborhood remain unaltered.

***Relative Location* SDCA model** In the *Relative Location* model, the transition functions all have access to the exact relative *location* and *state* of each neighbor site. Define a *neighbor* of site i, $n_i \in S \times Z^d$, as a pair that specifies the state (by a single label) and relative location of the neighboring site (as a d-tuple of coordinates). Let $W = S \times Z^d$ be the set of all possible neighbors, and $\mathcal{F}_W$ (called the *neighborhood function*) be the set of all possible finite, nonempty, partial functions that map Z^d to W. The local state transition function $\sigma : \mathcal{F} \to W$ maps neighborhood functions to the state set of SDCA. The local link transition function $\lambda : \mathcal{F} \times \mathcal{F} \times \{0, 1, 2\} \to \{0, 1\}$ maps pairs of neighborhood functions (that define the neighborhoods of two sites, i and j and a number that specifies the status of the link between i and j: value *zero* meaning that i and j are neither direct neighbors nor next-nearest neighbors; value *one* meaning that i and j are immediate neighbors; and value *two* meaning i and j are next-nearest neighbors) to one of two link states: *zero*, meaning no link between i and j, and *one*, meaning a link exists.

***Labeled Links* SDCA model** The *Labeled Links* model removes from M_{R}'s transition functions any dependency on the exact relative location of a site's neighbors, but allows the links to still be *labeled* so that the transitions functions can distinguish one link from another. This ability to "label" links paves the way for us to define SDCA with unidirectional links, since the labels can be used to distinguish between the *input* and *output* links to a site. Consider, for example, the UL-version of M_{L}. Labeling the links by natural numbers, $\mathcal{N}$, we define a neighbor of site i, as a pair (q, n), where $q \in S$ labels the state of the neighboring site, and $n \in \mathcal{N}$ labels the link between i and its neighbor. Site i is defined as the direct neighbor linked via the 0th link, and the set of all possible neighbors, $W = S \times \mathcal{N}$. As for M_{R}, $\mathcal{F}_W$ is the set of neighborhood functions that map Z^d to W, the local state transition function $\sigma : \mathcal{F} \to S$ maps neighborhood functions to states in S, and the local link transition function $\lambda : \mathcal{F} \times \mathcal{F} \times \{0, 1, 2\} \to \{0, 1\}$ maps pairs of neighborhood functions and a number to either the values *zero* (unlinked) or *one* (linked).

***Symmetric Links* SDCA model** The *Symmetric Links* model imposes the strictest constraint of all by doing away with all means by which the local transition functions may distinguish different neighborhood orientations. Consider the unbounded link version of M_{S}. Assume the SDCA has a total of n states, and let $S = \{1, 2, \ldots, n\}$. Let $\vec{n}_i \in N^n$ be an n-dimensional vector such that $(n_i)_k$ is equal to the number of site i's neighbors in state k. Then the local state transition function $\sigma : N^n \to S$ maps vectors in N^n to states in S, and the local link transition function $\lambda : N^n \times \{0, 1, 2\} \to \{0, 1\}$ maps a vector in N^n and a link status label to either the values *zero* (unlinked) or *one* (linked). M_{S} can also be modified slightly to allow the local transition functions to retain knowledge of the state of site i: simply let $\sigma : S \times N^n \to S$ map the pair consisting of the *state* of site i and a vector that defines the distribution of states among i's immediate neighbors (excluding i). The local link function likewise assumes a similar form $\lambda\colon S^2 \times N^n \times \{0,1,2\} \rightarrow \{0,1\}$, where the first component of the 3-tuple input to *lambda* is a pair that defines the states of the two sites to which the link function is being applied.

SDCA as CA Simulators

What does it mean to say that one dynamical system *simulates* another? Heuristically, it means that, for certain initial states, one system behaves just like another [32,74]. Suppose we have two CA systems – CA and CA′ – defined by rules ϕ and ϕ', and initial states $\vec{\sigma} \in \Sigma$ and $\vec{\sigma}' \in \Sigma$, respectively. Then, loosely speaking, T iterations of CA are said to be "simulated" by nT $(n \geq 1)$ iterations of CA′, provided there exists some invertible function, $f\colon \Sigma \rightarrow \Sigma$, by which $\vec{\sigma}'$ is replaced by $f(\vec{\sigma})$. Simulation is a transitive relationship: if system B simulates system A, and another system C simulates B, then C also simulates A.

For example, a single site with a particular value in CA may be simulated by a fixed *block of sites* in CA′. After n steps, the blocks in CA′ evolve to exactly the same final state as the single time-step evolution of individual sites in CA. As a concrete example, consider the elementary (one-dimensional, binary valued, conventional CA) rules ϕ_{18} and ϕ_{90}:

	111	110	101	100	011	010	001	000
	↓	↓	↓	↓	↓	↓	↓	↓
ϕ_{18}:	0	0	0	1	0	0	1	0
ϕ_{90}:	0	1	0	1	1	0	1	0

Provided that *two time steps* under ϕ_{18} are carried out for every time step of rule ϕ_{90}, it is easy to show that under the block transforms $0 \rightarrow f(0) = 00$ and $1 \rightarrow f(1) = 10$, the evolution of arbitrary starting configurations under ϕ_{90} is reproduced – or *simulated* – by ϕ_{18}. For example, the global state $\vec{\sigma} =$ '0011000' – which evolves into $\phi_{90}(\vec{\sigma}) =$ '0111100' under ϕ_{90} – yields the same state (after it is block-transformed) that results from two iterations of ϕ_{18} applied to ϕ_{90}'s block-transformed initial state, $f(\vec{\sigma}) =$ '00001010000000':

$$\phi_{18}[\phi_{18}[f(\vec{\sigma})]] = 00101010100000 = f[\phi_{90}(\vec{\sigma})]. \quad (23)$$

Now consider the specific case of SDCA simulating a conventional CA (we follow Majercik [39]). First, because SDCA cannot be expected to preserve the local topology of a simulated CA, it is necessary to define separate *encoding* $(= e)$ and *decoding* $(= d)$ functions – $e\colon \Sigma_{\mathrm{CA}} \rightarrow \Sigma_{\mathrm{SDCA}}$ transforms the initial configuration of the CA systems to configurations in the SDCA system being used to simulate it (where Σ_{CA} and Σ_{SDCA} are the configurations spaces of CA and SDCA, respectively); and $d\colon \Sigma_{\mathrm{SDCA}} \rightarrow \Sigma_{\mathrm{CA}}$ effectively performs the inverse transformation. Encoding (and decoding) functions are called *structurally defined* if they are recursive and use a finite amount of information to encode (or decode) a given configuration; and are otherwise expected to transform quiescent states to quiescent states. Majercik further assumes that (1) e has access to the rule table of the conventional CA system being simulated; (2) d does not have access to the rule tables of either system; and (3) e and d must together satisfy the relation: $e \times d = \mathrm{Identity}(\Sigma_{\mathrm{CA}})$.

Denoting the global transition functions of the CA and SDCA systems by Φ_{CA} and Φ_{SDCA}, respectively, Φ_{SDCA} is said to *simulate* Φ_{CA} if there exist $m \geq 1, n \geq 1$ and structurally-defined functions $e\colon \Sigma_{\mathrm{CA}} \rightarrow \Sigma_{\mathrm{SDCA}}$ and $d\colon \Sigma_{\mathrm{SDCA}} \rightarrow \Sigma_{\mathrm{CA}}$, such that for any configuration $\vec{\sigma} \in \Sigma_{CA}$ and any $k \geq 1$,

$$\Phi^{kn}_{\mathrm{CA}}(\vec{\sigma}) = d\left[\Phi^{km}_{\mathrm{SDCA}}\left[e\left(\vec{\sigma}\right)\right]\right]. \quad (24)$$

If $m > n$ then Φ_{SDCA} simulates Φ_{CA} with a *slowdown* factor of m/n. If $m < n$ then Φ_{SDCA} simulates Φ_{CA} with a *speedup* factor of n/m.

SDCA Hierarchy of Models

Majercik [39] uses the three generalized models introduced above $(M_{\mathrm{R}}, M_{\mathrm{L}}$, and $M_{\mathrm{R}})$ to define a hierarchy of eight SDCA models of computation. At the top of his hierarchy (arranged from top-to-bottom in roughly, but not completely, decreasing order of computational strength; see discussion that follows) are the UL and BL versions of M_{R}: $\mathrm{SDCA}^{(8)}$ and $\mathrm{SDCA}^{(7)}$, respectively; followed by $\mathrm{SDCA}^{(6)} =$ UL version of M_{L}; $\mathrm{SDCA}^{(5)} =$ BL version of M_{L}; $\mathrm{SDCA}^{(4)} =$ a *finite labels* version of M_{L}; $\mathrm{SDCA}^{(3)} =$ UL version of M_{S}; $\mathrm{SDCA}^{(2)} =$ BL version of M_{S}; and, sitting on the lowest level (computationally speaking), is $\mathrm{SDCA}^{(1)} = \mathrm{SDCA}_0$.

A little thought suffices to establish certain relationships among the various classes. Give two classes, C_1 and C_2, let $C_1 \leq_{\mathrm{s}} C_2$ denote the fact that if, given any SDCA $S_1 \in C_1$ there exists an SDCA $S_2 \in C_2$ that simulates S_1. Then, for example, since any BL SDCA can be simulated by an unbounded links version of the same system, and a finite links version of M_{L} can be simulated by a bounded links version, we know immediately that $\mathrm{SDCA}^{(7)} \leq_{\mathrm{s}} \mathrm{SDCA}^{(8)}$, $\mathrm{SDCA}^{(4)} \leq_{\mathrm{s}} \mathrm{SDCA}^{(5)} \leq_{\mathrm{s}} \mathrm{SDCA}^{(6)}$, and $\mathrm{SDCA}^{(2)} \leq_{\mathrm{s}} \mathrm{SDCA}^{(3)}$. Similar reasoning [39] leads to the general relationship:

$$\begin{cases} \mathrm{SDCA}^{(3)} \leq_{\mathrm{s}} \mathrm{SDCA}^{(8)} \leq_{\mathrm{s}} \mathrm{SDCA}^{(6)}\,, & \text{and} \\ \mathrm{SDCA}^{(2)} \leq_{\mathrm{s}} \mathrm{SDCA}^{(7)} \leq_{\mathrm{s}} \mathrm{SDCA}^{(5)}\,. \end{cases} \quad (25)$$

Finally, since the unbounded links version of M_R has all the information necessary to construct the neighborhood partitions used by SDCA$_0$, and since SDCA$^{(8)}$ $\leq_\mathrm{s}$ SDCA$^{(6)}$, we see that SDCA$_0$ $\leq_\mathrm{s}$ SDCA$^{(6)}$ and SDCA$_0$ $\leq_\mathrm{s}$ SDCA$^{(8)}$.

Majercik's two main results, which we state without proof, are:

> *Majercik Theorem 1:* Given an arbitrary 1-dimensional conventional CA with radius $r = 1$, there exists an unbounded links version of M_R (= SDCA$^{(8)}$ of the SDCA hierarchy) that can simulate it with a speedup factor of two.

> *Majercik Theorem 2:* There exists a 1-dimensional finite links version of M_L (= SDCA$^{(4)}$) that can simulate an arbitrary k-state 1-dimensional conventional CA with radius $r = 1$ with a slowdown factor $O\left(k^{2r}\sqrt{2r\log k}\right)$.

Detailed proofs of these two theorems appear in [39] (where they are called *Theorems 4.4* and *4.5*, respectively). In chapter 5 of his thesis [39], Majercik presents an explicit construction of a CA-universal SDCA$^{(4)}$ computational model, and compares it to Albert and Culik's [3] construction of a 1-dimensional CA-universal conventional CA that simulates any 1-dimensional, k-state, radius r CA with an $O\left(k^{8r}\right)$ slowdown. Although Majercik's CA-universal SDCA uses more states than Albert and Culik's universal CA, it is also markedly faster.

The reason why the SDCA is faster is at least intuitively clear. An SDCA's dynamic links effectively endow an otherwise conventional CA with a random access memory. Since SDCA can establish links between any two sites a distance d apart in $O(\log d)$ time, *any* site potentially has access to the state of any other site. While it may be argued that sites in conventional CA can also access the states of other cells, they *cannot do so permanently*. Once information is accessed once and used, the connection is lost, and must subsequently be re-established. Moreover, the links in SDCA can potentially connect sites that are arbitrarily far apart; so that, once a small number of links are dynamically created, they continue to provide long-range communication channels throughout the network. Since the propagation of information in a conventional CA is necessarily limited in being able to flow one site at a time, the overall computational speed is obviously limited.

However, it is worth pointing out that while the computational strength of Majercik's CA-universal SDCA model undoubtedly derives from its ability to forge long-range communication links, the results as quoted from [39] do not tap into what is potentially SDCA's greatest strength; namely, the ability to *adaptively create links, even as a given computation unfolds*. In Majercik's model, the links are dynamically coupled to an actual computation only insofar as they are initially fixed as a function of the initial state. While the local structure certainly *evolves* (as it does in all SDCA systems, as the computation itself unfolds), it does so purely as a consequence of the SDCA rules, and not adaptively to the evolution.

Majercik concludes his thesis by speculating on how an *adaptive* variant of his CA-universal SDCA may be used to explore certain aspects of evolutionary learning. (Working from a different set of assumptions, Halpern [23,24] applies evolutionary programming techniques to SDCA$_0$ to explore what happens when the structure is allowed to play an *explicit* dynamic role in the computation; see next section.) The question of whether there exist SDCA-universal SDCA models – that are able to simulate certain classes of the SDCA hierarchy, for example – remains open.

SDCA & Genetic Algorithms

Genetic algorithms (abbreviated, GA) are a class of heuristic search algorithms and computational models of adaptation and evolution based on natural selection. In nature, the search for beneficial adaptations to a continually changing environment (i. e., evolution) is fostered by the cumulative evolutionary knowledge that each species possesses of its forebears. This knowledge, which is encoded in the chromosomes of each member of a species, is passed on from one generation to the next by a mating process in which the chromosomes of "parents" produce "offspring" chromosomes. A comprehensive review of GA is given by Mitchell [45].

While GAs may be effectively used to search for "interesting" topological structures (but for which the structures themselves do not play any dynamic role; see, for example, Lehmann [37]), Halpern [23] is the first to explore a novel hybrid algorithm between GA and SDCA, in which SDCA rules are used to *evolve a GA*. Weinert et al. [72] explore a related "structurally dynamic" GA model, in which links between adjacent individuals of a population are dynamically chosen according to deterministic or probabilistic rules. In this section, we follow Halpern [23,24].

Formally, GAs are defined by (1) an ensemble of "candidate solution" vectors, $\{\vec{s}_i\}$: $\vec{s}_i \in M_\mathcal{P} \subseteq R^n$, where M is the set of all possible solutions to a given "problem" $\mathcal{P}$ (the $\vec{s}_i$ are usually, but not always, defined as a string of binary numbers [45]), and (2) a "fitness function", $f(\vec{s})$, that represents how well a given $\vec{s}$ "solves" $\mathcal{P}$. The goal of the GA is to find the global optimal solution,

$\vec{s}^*$ such that (from the point of view of *maximizing* fitness, $f(\vec{s}) \leq f(\vec{s}^*) \equiv f^*, \forall \vec{s} \in M$. Optimization proceeds through the combined processes of *selection*, *breeding*, *mutation*, and *crossover replacement* [45]; to which – in the hybrid SDCA↔GA algorithm, Halpern adds the new feature of *self-selective neighborhood structure*.

It should be immediately noted that this is not an ad-hoc addition. Muhlenbein [47] points out that if each generation of a GA searches over the entire possible solution space, the algorithm may – depending on the fitness function – converge prematurely to a sub-optimal solution. To reduce the likelihood of this happening, Muhlenbein introduces a spatial population structure; restricting fitness and mating to neighborhoods called *demes*. Demes are geographically separate subpopulations in which candidate solutions evolve along disparate trajectories; though occasional mixing still occurs through the process of migration.

In Halpern's variant [23], an otherwise conventional GA is placed within the structure of SDCA$_0$ (i. e., the basic model defined by Eqs. (11)–(14)). Heuristically, this allows each candidate solution to "choose a community" with which to mate, during each generation. The choice of neighborhoods thus becomes an integral component of the GA, and is determined dynamically by the evolving solutions.

Halpern's algorithm proceeds as follows [24]: *(Step 1)* an initially random lattice (defined by adjacency matrix $l_{ij}^{(t=0)}$) is seeded with single-chromosome candidate solutions of fixed length, one per site; *(Step 2)* a fitness function, $f_i = \sum_{i=1}^{N} \delta_{ij}$, is defined to assign a numerical measure of "optimality" to each site (N is the number of sites, and δ_{ij} is the value – equal to 0 or 1 – of the jth gene of the ith chromosome; *(Step 3)* each site i ranks each of its nearest and *next*-nearest neighbors according to f_i; *(Step 4)* each site disconnects with a fraction, f_D, of its least-fit neighbors, and connects with a fraction, f_C, of its fittest next-nearest neighbors; *(Step 5)* each site randomly mates with one of its nearest neighbors (i. e., the usual processes of mutation and crossover operations are applied [45]); *(Step 6)* the least fit members of the population are replaced by the offsprings from *Step* 5; and *(Step 7)* loop through steps *5–7*, until some suitable "optimality" threshold (or some other convergence criterion) is satisfied.

Halpern [23,24] reports a wide range of resulting behaviors, collectively suggesting a clear relationship between the parameters defining the GA optimization and lattice connectivity. Of particular interest are the dynamic conditions for which the fitness-based creation and deletion of links increases the rate of growth of overall fitness. The fastest convergence occurs when lattice connectivity first increases, then decreases, then eventually levels off. In the first stage, the fittest possible communities are first established; in the second stage, connections with poorer candidate solution are deleted; finally, in the third stage, the system essentially "fine-tunes" its optimal solutions. Halpern [24] finds two different evolutionary paths toward high connectivity: (1) monotonic growth over time (for low mutation rates, p_μ), and (2) a *phase transition* between low and high degrees of connectivity (for some p_μ^*). Using SDCA↔GA hybrid model parameters $N = 100$ and $f_D = f_C = 0.1$, $p_\mu^* \approx 0.05$, for which Halpern [24] finds a sharp increase in the number of links per site between generations 350 and 450.

Despite the novelty of the approach, and the promising link between optimization rates and dynamic structure established in [23], concrete applications of the algorithm – except for Weinert et al. [72] work on a related hybrid GA algorithm – have yet to be developed. One suggestion, from Halpern [24], is to use the SDCA↔GA hybrid model for finding "optimal" connectivity patterns in parallel computers. The search algorithm may be used to directly model how component processors are connected, and decide to keep or sever existing links, or establish new ones, adaptively as a function of local fitness criteria.

Generalized SDCA Models

Despite SDCA being obviously more "complex" than conventional CA (and certainly more complex to formally define, if only because one must specify *both* σ *and* ℓ rules), the SDCA model nonetheless has more in common with *elementary* CA than with any of its brethren's more "complicated" variants. By "elementary" CA we mean the simplest one-dimensional CA with $\sigma \in \{0, 1\}$ and local neighborhoods consisting only of left and right (i. e., *nearest*) neighbors. Just as there are many generalizations of elementary CA – for example, increasing the state space to include σ's that take on one of N values, larger-sized neighborhoods, and memory, among many other possibilities – so too there are natural extensions of basic SDCA. In this section we discuss three generalizations: (1) rules that are *reversible in time*, (2) rules that retain a *memory of past states*, and (3) *probabilistic* rules.

Reversible SDCA

The first generalization of the basic SDCA model, explored extensively by Alonso-Sanz [7], is to apply the *Fredkin reversible-rule construction* to ℓ rules to render them reversible in time. Consider a conventional CA system that is first-order in time, $\sigma_i^{t+1} = \phi[\sigma_j^t \in \mathcal{N}_i]$, where $\mathcal{N}_i$ is the neighborhood around site i and, generally, $\sigma_i \in \mathcal{Z}_k$. The *Fredkin* construction converts this system into an explicitly invertible one that is *second-order* in time by subtracting the value of the center site at time $t-1$:

$$\sigma_i^{t+1} = \phi\left[\sigma_j^t \in \mathcal{N}_i\right] \ominus_k \sigma_i^{t-1}, \tag{26}$$

where '$\ominus_k$' is subtraction *modulo-k*. Since Eq. (26) can be trivially solved for σ_i^{t-1} $\left(=\phi[\sigma_j^t \in \mathcal{N}_i] \ominus_k \sigma_i^{t+1}\right)$, we see that any *pair of consecutive configurations uniquely specifies the backwards trajectory of the system*. Moreover, this is true for arbitrary (and, in particular, *irreversible*) functions ϕ.

Now, exactly the same procedure may be applied to link functions:

$$\begin{cases} l_{ij}^{t+1} = \psi\left(\{\sigma_k^t\}, \{l_{ij}^t\}\right) \ominus_2 l_{ij}^{t-1}, \\ l_{ij}^{t+1} = \omega\left(\{\sigma_k^t\}, \{l_{ij}^t\}\right) \ominus_2 l_{ij}^{t-1}, \end{cases} \tag{27}$$

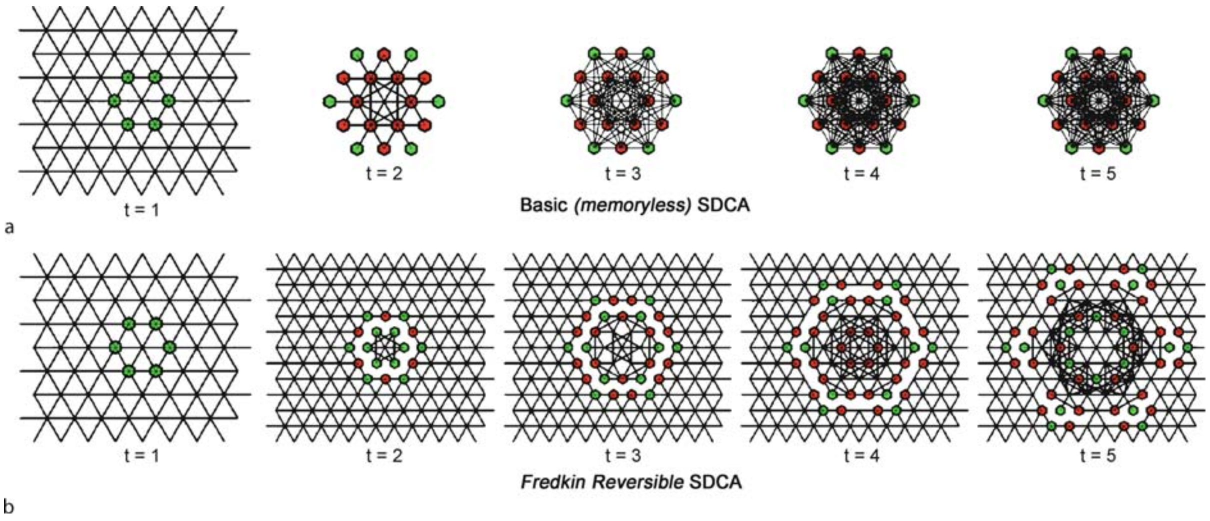
where $\ominus_2$ is subtraction *modulo*-2 (since links are obviously binary valued).

Following Alonso-Sanz [7,8], we consider these two specific SDCA link rules (which will also be used in a later example):

$$\begin{cases} \psi\left(\sigma_i^t, \sigma_j^t, l_{ij}^t\right) = 0 \text{ iff } l_{ij}^t = 1 \text{ and } \sigma_i^t + \sigma_j^t = 0\,, \\ \omega\left(\sigma_i^t, \sigma_j^t, l_{ij}^t\right) = 1 \text{ iff } l_{ij}^t = 0, \sigma_i^t > 0, \sigma_j^t > 0, \text{ and } D_{ij} = 2. \end{cases} \tag{28}$$

Figure 10 compares the evolution of the *Fredkin reversible* version of these rules to their memoryless counterpart. Both evolutions start on a two dimensional hexagonal lattice, and values evolve according to the three-state (i. e., $\sigma \in \{0,1,2\}$), next-nearest neighborhood **T** *beehive* rule. The *beehive* rule is defined explicitly by assigning one of three values $(0, 1,$ or $2)$, to each possible 3-tuple, (N_0, N_1, N_2), that gives the number of local sites with N_0 0s, N_1 1s, and N_2 2s [7]: $(0,0,6) \to 0$, $(0,1,5) \to 1$, $(0,2,4) \to 2$, $(0,3,3) \to 1$, $(0,4,2) \to 2$, $(0,5,1) \to 0$, $(0,6,0) \to 0$, $(1,0,5) \to 0$, $(1,1,4) \to 2$, $(1,2,3) \to 2$, $(1,3,2) \to 2$, $(1,4,1) \to 1$, $(1,5,0) \to 1$, $(2,0,4) \to 0$, $(2,1,3) \to 0$, $(2,2,2) \to 2$, $(2,3,1) \to 2$, $(2,4,0) \to 0$, $(3,0,3) \to 0$, $(3,1,2) \to 2$, $(3,2,1) \to 2$, $(3,3,0) \to 0$, $(4,0,2) \to 0$, $(4,1,1) \to 0$, $(4,2,0) \to 2$, $(5,0,1) \to 2$, $(5,1,0) \to 0$, $(6,0,0) \to 0$.

The top row of Fig. 10 shows the first four steps ($t = 1, 2, 3,$ and 4) in the memoryless evolution of the initial "ring" of sites that appears at $t = 1$. The link rules used for this run are those defined in Eq. (28). Since the decoupler removes links between pairs of sites whose values are equal to zero, most of the lattice disappears after a single time step, and both value and link activity is confined to a small region. After two more steps of changes, the system quickly attains a fixed point: $\{\sigma^t, \ell_{ij}^t\} = \{\sigma^{t=4}, \ell_{ij}^{t=4}\}$ for all $t \geq 5$. While the frequency of states is not constrained to total six for a dynamic lattice, the *beehive* rule is unchanged; if the sum of frequencies at a given site exceeds six, the site value remains the same.



Structurally Dynamic Cellular Automata, Figure 10
Comparison between first few time steps of a a *memoryless* SDCA, evolving according to link rules defined in Eq. (28), and b the *Fredkin reversible* versions of these rules (obtained by applying Eq. (27) to Eq. (28)). In both cases, σ's evolve according to the *beehive* rule defined in the text. (Reproduced with permission from [7])

The bottom row shows the evolution of the *Fredkin reversible* versions of the rules defined in Eq. (28) (to simplify the visualization, links along the border sites are not shown). In contrast to the basic SDCA version, the initial lattice in this case *does not* decay. Since, according to Eq. (27) (which assumes that $\ell_{ij}^{t=0} = \ell_{ij}^{t=1}$), the initial hexagonal lattice is subtracted from the evolved structure at $t = 1$(*modulo*-2), the original graph is effectively restored, and the outlying regions appear undisturbed.

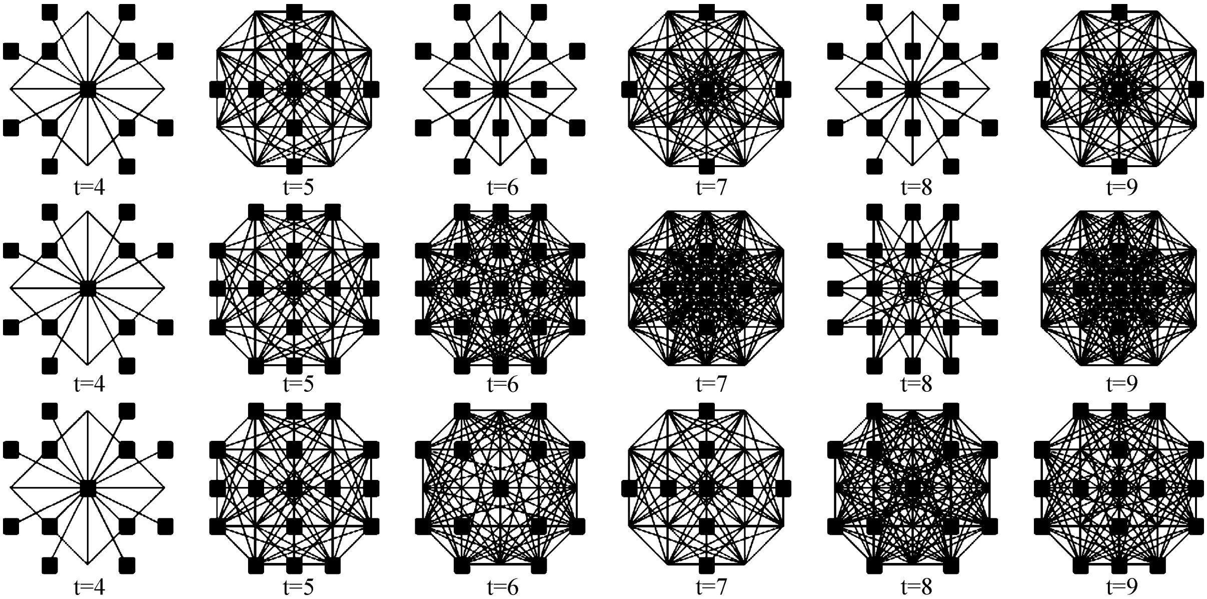
SDCA with Memory

A second generalization to the basic SDCA model, introduced and studied by Alonso-Sanz and Martín [6,7,8], is to endow both σ-rules and ℓ-rules with *memory*. The rules for conventional memoryless CA and SDCA, depend only on neighborhood configurations that appear on the immediately preceding time step. Therefore, rules may be said to possess a "memory" of depth m if they depend explicitly on values (in the case of CA), or on both values and link states (in the case of SDCA), that existed on m previous time steps. We note, in passing, that since the *Fredkin construction* couples states at times $t+1$, t and $t-1$, reversibility may be considered a specific form of memory that extends backwards a single step.

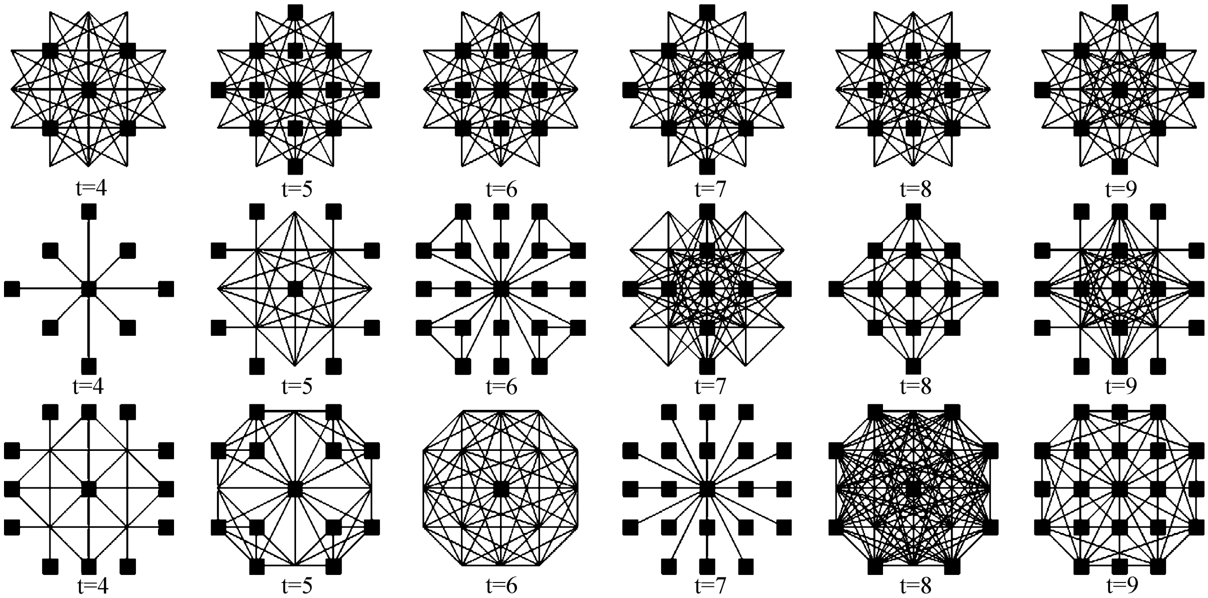
Of course, there is no unique prescription for introducing a dependency on past values; and a variety of alternative memory mechanisms have been proposed in the literature (for example, see page 43 in [32] and page 118 in [74]). We focus our discussion on the approach proposed by Alonso-Sanz (14), and for the moment confine our attention to value rules, $\phi: \sigma \to \sigma'$. Alonso-Sanz's approach is to preserve the form of the transition rule, but have it act on an effective site value that is a weighted function of its m prior values.

This is done by introducing a memory-endowed value rule, ϕ_m, that – in contrast to its memoryless version, ϕ – is not, in general, a function of a given site's current value, σ_i, *alone*, but is instead a function of the transformed value, $s = \mathcal{M}_\phi(\sigma; m, \alpha)$, obtained from σ_i's past m values: $\phi_m: s \to \sigma'$, where $0 \le \alpha \le 1$ is a numerical *memory factor*. The value transforming memory function, $\mathcal{M}$, assumes the following specific form (to avoid confusion, note that in Eqs. (29) and (30), σ_i^x means *the value of* σ_i *at time* $t = x$, and α^x means *the numerical quantity* α *raised to the power* x):

$$s_i^t = M_\phi\left(\sigma_i^t; m, \alpha\right) = \begin{cases} 1 & \text{if } \left(\hat{\sigma}_i^t\right)_m > 1/2, \\ \sigma_i^t & \text{if } \left(\hat{\sigma}_i^t\right)_m = 1/2, \\ 0 & \text{if } \left(\hat{\sigma}_i^t\right)_m < 1/2, \end{cases} \qquad (29)$$



Structurally Dynamic Cellular Automata, Figure 11
Sample runs of a SDCA with memory for memory weighting $\alpha = 0.6$. The SDCA is initialized as a Euclidean four-neighbor lattice, and evolves according to the *parity* T σ-rule and the two ℓ rules defined in Eq. (28). The first row of evolving patterns applies memory only to *values*; the second row applies memory only to *links*, and the third row shows the evolution when memory is applied to both. (Reproduced by permission from [8])



Structurally Dynamic Cellular Automata, Figure 12
Sample runs of the same SDCA shown in Fig. 11, but with memory weighting $\alpha = 1.0$. (Reproduced by permission from [8])

where

$$\left(\hat{\sigma}_i^t\right)_m = \frac{\sigma_i^t + \sum_{\Delta t=1}^m \alpha^{\Delta t} \cdot \sigma_i^{t-\Delta t}}{1 + \sum_{\Delta t=1}^m \alpha^{\Delta t}} \,. \tag{30}$$

At any given time, t, the depth m can never exceed $t-1$. Our discussion follows Alonso-Sanz [8], and sets $m(t) \equiv t-1$ for all t; i. e., we assume that $\mathcal{M}_\phi(\sigma; m, \alpha)$ yields a weighted mean value of *all* the previous values of a given site. In practice, memory becomes active only after a certain number of initialization steps, here taken to be three; with seeded values $s_i^1 = \sigma_i^1$ and $s_i^2 = \sigma_i^2$.

Memory can be added to link rules in a similar manner. The form of the link rules (ψ and ω) remains the same, but rather than acting on a graph that is defined by its adjacency matrix, ℓ_{ij}^t, ψ and ω instead act on the memory-transformed values, $L = \mathcal{M}_{(\psi,\omega)}(\ell; m, \alpha)$:

$$L_{ij}^t = M_{(\psi,\omega)}\left(l_{ij}^t; m, \alpha\right) = \begin{cases} 1 & \text{if } \left(\hat{l}_{ij}^t\right)_m > 1/2, \\ l_{ij}^t & \text{if } \left(\hat{l}_{ij}^t\right)_m = 1/2, \\ 0 & \text{if } \left(\hat{l}_{ij}^t\right)_m < 1/2, \end{cases} \tag{31}$$

where

$$\left(\hat{l}_{ij}^t\right)_m = \frac{l_{ij}^t + \sum_{\Delta t=1}^m \alpha^{\Delta t} \cdot l_{ij}^{t-\Delta t}}{1 + \sum_{\Delta t=1}^m \alpha^{\Delta t}} \,. \tag{32}$$

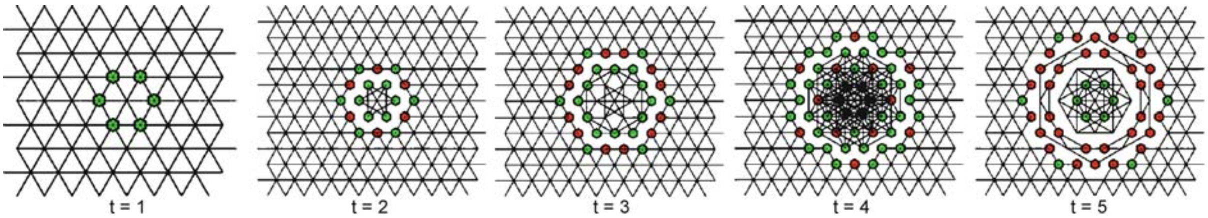
As for memory-endowed σ-rules, the memory for link rules is activated only on the third iteration step, and the system is initialized by setting $L_i^1 = \sigma_i^1$ and $L_i^2 = \sigma_i^2$.

Figures 11 and 12 show the effects of applying partial memory weighting ($\alpha = 0.6$) and full memory ($\alpha = 0.6$), respectively, to a SDCA that starts with a Euclidean four-neighbor lattice, and evolves according to the *parity* **T** σ-rule (that assigns a value *zero* to a site if the sum of the values in its neighborhood is *even*, and assigns the value *one* if the sum is *odd*) and the ℓ rules defined above in Eq. (28). The first row of evolving patterns (for each α) applies memory only to *values*; the second applies memory only to *links*, and the third appliers memory to both. Figure 13 shows the reversible *beehive* SDCA shown in Fig. 10, but with full memory ($\alpha = 1.0$).

Probabilistic SDCA

Another natural extension of the basic SDCA model is to replace the set of explicit σ- and/or ℓ-rules with probabilities. In this way one can study the evolution of a system that undergoes random but σ-dependent lattice changes. For example, this may be useful for studying genetic networks in which new links are forged (with a given probability) only if both genes are active, and existing connections are broken if both sites are inactive.

Following Halpern and Caltagirone [22], consider the *parity* **T** σ-rule and the following probabilistic versions of



Structurally Dynamic Cellular Automata, Figure 13
Sample runs of a reversible *beehive* SDCA with full memory ($\alpha = 1.0$); compare to Fig. 10. (Reproduced by permission from [8])

decoupler (ψ_p) and coupler rules (ω_p):

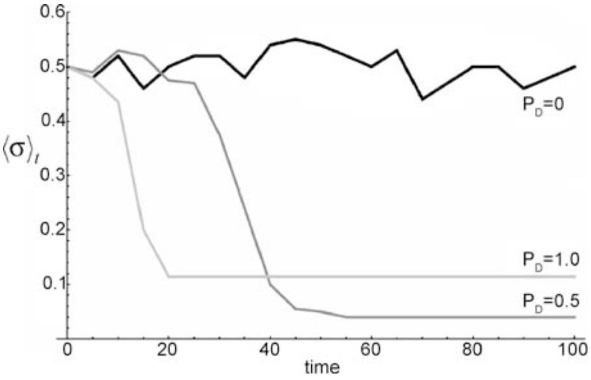
$$(decoupler): \begin{cases} l_{ij}^{t+1} = \psi_p\left(l_{ij}^t;\sigma_i^t,\sigma_j^t,p_D\right), \\ \psi_p \equiv 1-\delta\left(\sigma_i^t+\sigma_j^t,0\right)\cdot\delta\left(p_D>r\right), \end{cases}$$

$$(coupler): \begin{cases} l_{ij}^{t+1} = \omega_p\left(l_{ij}^t;\sigma_i^t,\sigma_j^t,p_D\right), \\ \omega_p \equiv \delta\left(D_{ij},2\right)\cdot\delta\left(\sigma_i^t+\sigma_j^t,2\right)\cdot\delta\left(p_C>r\right), \end{cases} \tag{33}$$

where P_D and P_C are the *decoupler* and *coupler* probabilities, respectively, and r is a random number between 0 and 1.

Thus, ψ_p unlinks two previously linked sites with probability P_D *if and only if* the sum of their site values is *zero*; and ω_p links two previously unlinked sites with probability P_C *if and only if* they are next-nearest neighbors and the sum of their site values is *two*.

Figure 14 shows time series plots of $\langle\sigma\rangle$ as a function of time for three different cases: (1) $P_D = 0$ (no decoupling at all), (2) $P_D = 1/2$, and (3) $P_D = 1$ (decoupler rule applied 100% of the time (consistent with *non*-probabilistic SDCA rules). We see that changing P_D induces qualitatively different σ behavior, that ranges from small fluctuations around $\langle\sigma\rangle \sim 0.5$ (for $P_D = 0$), to decay to small static values ($\langle\sigma\rangle = 0.05$ for $P_D = 1/2$, and $\langle\sigma\rangle = 0.12$ for $P_D = 1$).



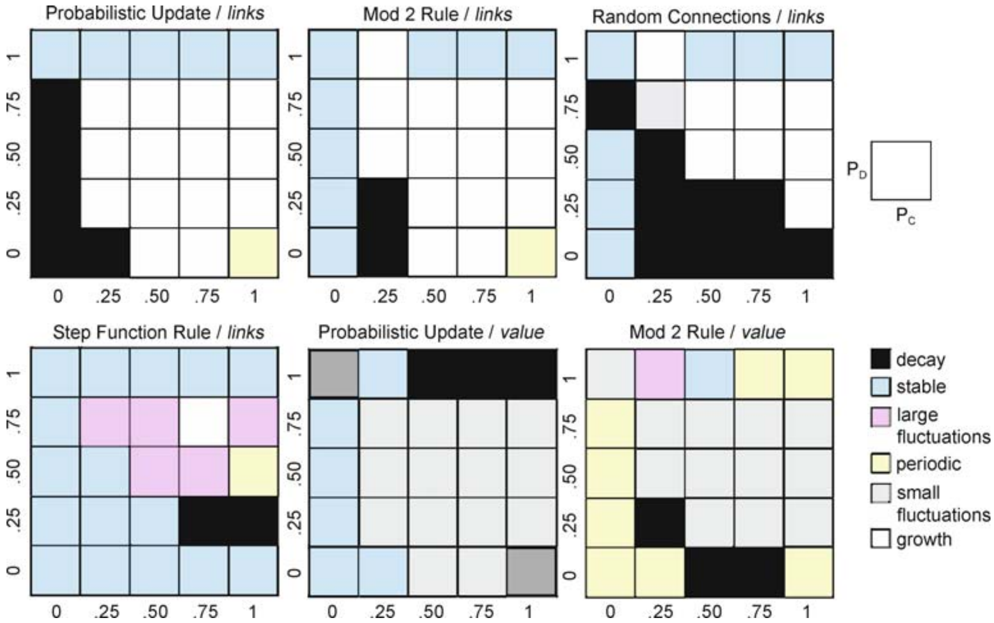
Structurally Dynamic Cellular Automata, Figure 14
Time series of average σ value, $\langle\sigma\rangle_t$, for the Halpern-Caltagirone rules (defined in Eq. (33)) and for three values of *decoupler* probability: $P_D = 0$, $P_D = 1/2$, and $P_D = 1$. (Reproduced with permission from [22])

Halpern and Caltagirone [22] have studied a wide range of probabilistic SDCA, using random initial σ configurations, *step-function*, *parity*, and Conway's *life* σ-rules, Cartesian and random initial lattice structures, and various probabilities $0 \leq P_D \leq 1$ and $0 \leq P_C \leq 1$. Some of their results are reproduced (with permission) in the *behavioral phase* plots shown in Fig. 15. (The *step-function* rule is defined by $\sigma_i^{t+1} = 0$ if and only if $\sum_j \ell_{ij}^t \sigma_i^t > 2$ and $\sigma_i^{t+1} = 1$ if and only if $\sum_j \ell_{ij}^t \sigma_i^t \leq 2$; Conway's *life* rule assigns $\sigma^{t+1} = 1$ to a site if and only if $\sigma^{(t)} = 0$ and the sum of values in its neighborhood at time t is equal to 3 or $\sigma^t = 1$ and the sum of values is equal to 2 or 3; otherwise $\sigma^{t+1} = 0$.)

Figure 15 shows a wide range of possible behaviors. Consider, for example, the number of *links per site* for the case where the lattice is updated with probabilistic ℓ-rules and the σ's are all random (shown at the top left of the figure). Four distinct classes of behavior appear, with *growth* dominant for most values of P_D and P_C.

Pure decoupling (or pure coupling) leads to complete decay (or growth to a stable state); a mixed state of coupling/decoupling generally yields slow growth. Periodic behavior occurs only for $P_D \sim P_C \sim 1$. Compare this behavior with the cases where the σ-rule is either the *parity* value rule (shown in the middle of the top row of Fig. 15) or the *step-function* rule (shown at left bottom of the figure). While the *parity* rule also displays four similar phases (growth to stability, decay to stability, incomplete growth, and incomplete decay), decaying structures eventually reach a stable (not null) final state. The *step-function* rule shows an even greater variety of possible behaviors, and appears more sensitive to small changes in link probabilities.

The probabilistic SDCA system discussed in this section adds a stochastic element *specifically to SDCA*. Of course, there are other ways of injecting stochasticity into a CA with dynamic topology. For example, Makowiec [40] combines the deterministic evolution of a conventional



Structurally Dynamic Cellular Automata, Figure 15

Behavioral phase **plots summarizing the long term evolution for several different σ and ℓ-rules defined in Eq. (33). The *x* and *y* axes for each plot depict values ($\in \{0, .25, .5, .75, 1\}$) of P_C and P_D, respectively. There are six classes of behavior: *growth*, *decay*, *stability*, large and small *fluctuations* (around a stable lattice), and *periodicity*. The initial graph is a Cartesian four-neighbor lattice in each case except for the *top-right* plot (labeled *Random connections/links*) for which the initial graph is random. (Reproduced with permission from [22])**

CA with an asynchronous stochastic evolution of its underlying lattice (patterned after the Barabasi–Albert [9] model of degree distributions in small-world networks), to explore the influence of dynamic topology on the zero-temperature limit of ferromagnetic transitions.

Random Dynamics Approximation

For cases in which the structure and value configurations are both sufficiently random and uncorrelated, a *random dynamics approximation* (abbreviated, RDA) may suffice to qualitatively predict how the system will tend to evolve under a specific rule set; for example, to predict whether a given rule is more (or less) likely to yield unbounded growth, to eventually settle into a low periodic state, or to simply decay. The idea is to approximate the real SDCA as a *mean-field*; that is, assume all local value and structural correlations are close to zero (and can thus be ignored), and replace all specific site values and local link geometries with average, or effective, values.

More precisely, assuming that (1) the probability $p_n^{(\sigma_i)}$ of a site '*i*' having value $\sigma = 1$ at time $t = n$ is the same for all sites – so that $p_n^{(\sigma_i)} = p_n^{(\sigma)}$ for all *i* – and (2) that the probability $p_n^{(l_{ij})}$ of two sites '*i*' and '*j*' being linked at $t = n$ is the same for all pairs of sites – so that $p_n^{(l_{ij})} = p_n^{(l)}$ for all *i* and *j* – the RDA evolution equations may be written formally as follows:

$$\begin{cases} p_{n+1}^{(\sigma)} = F_{\mathrm{RDA}}[p_n^{(\sigma)}, p_n^{(l)}; \Gamma_{\mathrm{SDCA}}], \\ p_{n+1}^{(l)} = G_{\mathrm{RDA}}[p_n^{(\sigma)}, p_n^{(l)}; \Gamma_{\mathrm{SDCA}}], \end{cases} \tag{34}$$

where SDCA's rule Γ_{SDCA} (defined in Eq. (14)) is included, formally, to remind us that the functional forms assumed by F_{RDA} and G_{RDA} will be different for different Γ_{SDCA}s.

The first function, F_{RDA}, is the easier of the two to calculate. For any given site with degree *d* we simply count the total number of ways to distribute the local σ-values among the *d* possible neighboring sites to obtain the desired sums that define a given rule. In this way we find the average expected σ density at $t = n + 1$, assuming all sites in the lattice have the same degree *d* at time $t = n$:

$$p_{n+1}^{(\sigma)}(d, p_n^{(\sigma)}) = \begin{cases} \sum_{\{\alpha\}} \binom{d+1}{\alpha} \left[p_n^{(\sigma)}\right]^{\alpha} \left(1 - p_n^{(\sigma)}\right)^{d+1-\alpha} & \leftrightarrow \mathbf{T} \\ \sum_{\{\alpha_0\}} \binom{d}{\alpha_0} \left[p_n^{(\sigma)}\right]^{\alpha_0+1} \left(1 - p_n^{(\sigma)}\right)^{d-\alpha_0} & \\ \quad + \sum_{\{\alpha_1\}} \binom{d+1}{\alpha_1} \left[p_n^{(\sigma)}\right]^{\alpha_1} \left(1 - p_n^{(\sigma)}\right)^{d+1-\alpha_1} & \leftrightarrow \mathbf{OT} \end{cases} \tag{35}$$

We then get $F_{\mathrm{RDA}} \to p_{n+1}^{(\sigma)} = \sum_d P(d; p_n^{(l)}) \cdot p_n^{(\sigma)}(d, p_n^{(\sigma)})$ as an average over all possible degrees, where $P(d; p_n^{(l)})$ is the probability that any site has exactly d neighbors. Since this means that, out of a total of $N-1$ possible neighbors, a given site must have exactly d links, and not be connected to any of the remaining $(N-1-d)$ sites, we have by inspection:

$$P\left(d; p_n^{(l)}\right) = \begin{pmatrix} N-1 \\ d \end{pmatrix} \left[p_n^{(l)}\right]^d \left(1 - p_n^{(l)}\right)^{N-1-d} . \quad (36)$$

To calculate the second function in Eq. (34) ($= G_{\mathrm{RDA}}$), we first define the local transition functions

$$\begin{cases} p_n^a\,(d_1, d_2, \lambda) \\ = \mathrm{Prob}\left(l = 1 \to l' = 0 \mid d_i = d_1, d_j = d_2, \left|A_{ij}\right| = \lambda\right), \\ p_n^b\,(d_1, d_2, \lambda) \\ = \mathrm{Prob}\left(D = 2 \to l' = 1 \mid d_i = d_1, d_j = d_2, \left|A_{ij}\right| = \lambda\right), \end{cases} \quad (37)$$

which give the probabilities that any two sites – i and j – will be *disconnected* (p_n^a) or *connected* (p_n^b) if they have prescribed degrees $d_i = d_1$ and $d_j = d_2$, and are each linked to the *same* λ sites in the shared neighbor set, A_{ij} (see Fig. 1). In the case of type-**T** σ- and ℓ-rules, p_n^a and p_n^b are given explicitly by (**OT** versions of σ- and ℓ-rules, and **RT** versions of ℓ-rules are defined by similar, but slightly more complicated, expressions):

$$\begin{cases} p_n^a\,(d_1, d_2, \lambda) = \sum_k \begin{pmatrix} d_1 + d_2 - \lambda \\ \beta_k \end{pmatrix} \left[p_n^{(\sigma)}\right]^{\beta_k} \\ \qquad \cdot \left(1 - p_n^{(\sigma)}\right)^{d_1 + d_2 - \lambda - \beta_k}, \\ p_n^b\,(d_1, d_2, \lambda) = \sum_k \begin{pmatrix} d_1 + d_2 + 2 - \lambda \\ \varepsilon_k \end{pmatrix} \left[p_n^{(\sigma)}\right]^{\varepsilon_k} \\ \qquad \cdot \left(1 - p_n^{(\sigma)}\right)^{d_1 + d_2 + 2 - \lambda - \varepsilon_k}, \end{cases} \quad (38)$$

where β_k and ϵ_k refer to the sums that appear in Eqs. (11) and (12). The total probability that any two sites will be disconnected ($l = 1 \to l' = 0$) or connected ($D = 2 \to l' = 1$) – P_n^a and P_n^b, respectively – may then be obtained by summing over all possible local topologies:

$$\begin{cases} P_n^a \equiv \left\langle \mathrm{Prob}\left(l = 1 \to l' = 0\right) \right\rangle \\ \quad = \sum_{d_1} \sum_{d_2} \sum_{\lambda} P_1\,(d_1, d_2, \lambda) \cdot p_n^a\,(d_1, d_2, \lambda), \\ P_n^b \equiv \left\langle \mathrm{Prob}\left(D = 2 \to l' = 1\right) \right\rangle \\ \quad = \sum_{d_1} \sum_{d_2} \sum_{\lambda} P_2\,(d_1, d_2, \lambda) \cdot p_n^b\,(d_1, d_2, \lambda), \end{cases} \quad (39)$$

where

$$\begin{cases} P_1\,(d_1, d_2, \lambda) = \mathrm{Prob}(\text{sites } i, j \mid l_{ij} = 1 \text{ have } d_i = d_1, \\ \qquad\qquad d_j = d_2, \left|A_{ij}\right| = \lambda), \\ P_2\,(d_1, d_2, \lambda) = \mathrm{Prob}(\text{sites } i, j \mid l_{ij} = 0 \text{ have } d_i = d_1, \\ \qquad\qquad d_j = d_2, \left|A_{ij}\right| = \lambda). \end{cases} \quad (40)$$

To find P_1 we need to count, from among the remaining $N-2$ sites, the number of ways of selecting disjoint sets S_1, containing $d_1 - 1 - \lambda$ sites linked only to i; S_2, consisting of $d_2 - 1 - \lambda$ sites connected only to j; and S_3, with λ sites linked to *both* i and j. But this is simply a multinomial coefficient, so we can write:

$$P_1\,(d_1, d_2, \lambda) = \frac{(N-2)_{d_1 + d_2 - \lambda - 2}}{(d_1 - 1 - \lambda)!\,(d_2 - 1 - \lambda)!\lambda!} \cdot \left[p^{(l)}\right]^{d_1 + d_2 - 2} \left(1 - p^{(l)}\right)^{2(N-1) - d_1 - d_2}, \quad (41)$$

where $(n)_k \equiv n(n-1)\cdots(n-k+1)$. Similarly, for P_2, we need to count the number of ways of choosing $d_1 - \lambda$ sites from i, $d_2 - \lambda$ sites from j, and λ sites from both:

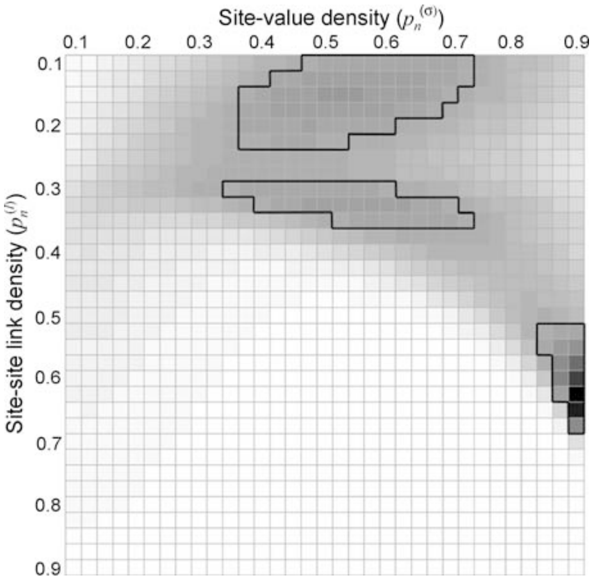
$$P_2\,(d_1, d_2, \lambda) = \frac{(N-3)_{d_1 + d_2 - \lambda - 2}}{(d_1 - \lambda)!\,(d_2 - \lambda)!\lambda!} \cdot \left[p^{(l)}\right]^{d_1 + d_2} \left(1 - p^{(l)}\right)^{2(N + \lambda - 3) - d_1 - d_2}. \quad (42)$$

The second (link-update) function of the pair of functions in Eq. (34) is thus given by

$$G_{\mathrm{RDA}} \to p_{n+1}^{(l)} = p_n^{(l)} \cdot \left(1 - P_n^a\right) + \left(1 - p_n^{(l)}\right) \cdot P_{D=2} \cdot P_n^a, \quad (43)$$

where, assuming that two sites, i and j, are not themselves connected, $P_{D=2}$ = probability that there exists at least one site k, such that $D_{ik} = D_{jk} = 1$, which implies that $P_{D=2} = 1 - \mathrm{Prob}\,(\text{there is no such } k) = 1 - (1 - [p_n^{(l)}]^2)^{N-2}$; and P_1 and P_2 are defined in Eqs. (41) and (42).

A *structural equilibrium* is established when $p_{n+1}^{(l)} \approx p_n^{(l)}$, which happens when the average number of new connections is equal to the average number of link deletions: $P_n^b \cdot \langle N_{\mathrm{nn}} \rangle = P_n^a \cdot \langle \mathrm{deg} \rangle$, where $\langle \mathrm{deg} \rangle = p_n^{(l)}(N-1)$ is the average *degree*, and $\langle N_{\mathrm{nn}} \rangle$ is the average number of *next-nearest neighbors*. For SDCA rules that naturally tend to produce graphs with minimal site value and structural



Structurally Dynamic Cellular Automata, Figure 16
Density-plot of $\gamma_{\text{c::d}}$ for an OT *decoupler* rule: $\{(0,0),(1,1),(1,2),(2,2)\}$; an OT *coupler* rule: $\{(1,1)\}$; $0.1 \leq p_n^{(\sigma)} \leq 0.9$; and $0.1 \leq p_n^{(l)} \leq 0.9$; the rectangular area highlighted in black denotes the "equilibrium boundary" that separates regions of growth and decay

correlations, the predicted ratio of RDA link creations to deletions, $\gamma_{c:d} \equiv P_n^b \cdot \langle N_{\text{nn}} \rangle / P_n^a \cdot \langle \text{deg} \rangle$, may be used to predict qualitatively how the graphs will evolve. Since the average number of pairs of sites a distance $D = 2$ apart $= \binom{N}{2} \cdot P_{D=2} = \langle N_{\text{nn}} \rangle \cdot N/2$, we find that:

$$\gamma_{\text{c::d}} = \frac{P_n^b}{P_n^a} \cdot \frac{\left(1 - p_n^{(l)}\right)}{p_n^{(l)}} \cdot \left\{ 1 - \left(1 - \left[p_n^{(l)}\right]^2\right)^{N-2} \right\} . \quad (44)$$

$\gamma_{\text{c::d}}$ is also implicitly a function of site-value density, since $p_n^{(\sigma)}$ appears in both P_n^a and P_n^b, defined in Eq. (39).

Figure 16 shows a grayscale density-plot of $\gamma_{\text{c::d}}$ for an **OT** *decoupler* rule: $\{(0,0),(1,1),(1,2),(2,2)\}$; an **OT** *coupler* rule: $\{(1,1)\}$; $0.1 \leq p_n^{(\sigma)} \leq 0.9$; and $0.1 \leq p_n^{(l)} \leq 0.9$. Areas that are close to *white* represent combinations of $(p_n^{(\sigma)}, p_n^{(l)})$ for which $\gamma_{\text{c::d}} \ll 1$, and which therefore predict "decay"; areas that are close to *black* represent combinations of $(p_n^{(\sigma)}, p_n^{(l)})$ for which $\gamma_{\text{c::d}} \gg 1$, and predict "growth"; the rectangular area highlighted in black denotes the "equilibrium boundary" that separates regions of growth and decay.

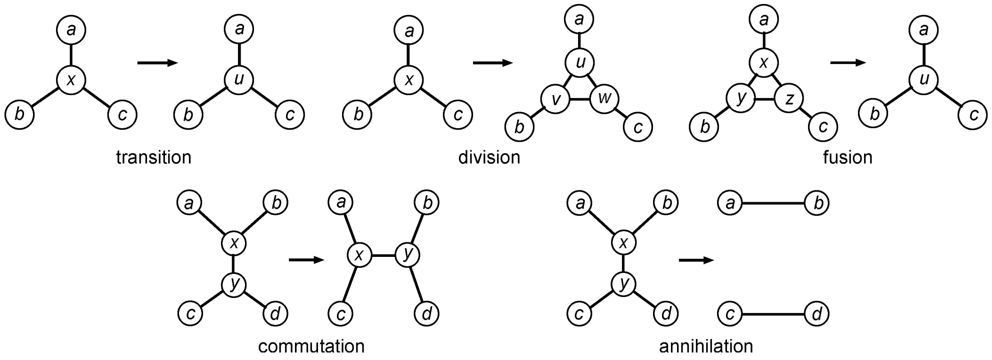
Related Graph Dynamical Systems

The original SDCA model [28] represents *one* (albeit not entirely arbitrary) approach to dynamically coupling site values ($\{\sigma_i\}$) and topology ($\{l_{ij}\}$), of the normally quiescent lattice. Since this model was primarily introduced as a general tool to explore self-organized emergent *geometries*, σ values are an integral dynamic component only because SDCA's original rules were conceived to generalize conventional CA rules, not replace them. Moreover, SDCA's link rules are, by design, close analogs of their conventional-CA brethren; this is the reason why SDCA's ψ and ω rules assume the familiar **T** and **OT** (and related **RT**) forms, as defined in Sect. "The Basic Model". Indeed, while the preceding sections of this article have introduced several generalizations – such as the addition of probabilistic rules, reversibility and memory – in each case, the basic *form* of the rules (as defined in Eqs. (11), (12), and (13)) has remained essentially the same. However, just as for conventional CA, an almost endless variety of different *kinds* of rules can in principle be defined; including rules that alter the geometry but are *not* functions of the σ states. In this section, we look at two illustrative examples of SDCA-like dynamical systems: one that uses coupled σ-ℓ rules, and another whose rules depend only on *topology*.

Graph Rewriting Automata

Tomita, Kurokawa, and Murata [67,68,69,70,71] have recently introduced *graph rewriting automata* (abbreviated, GRA), in which both links and (the number of) sites are allowed to change. Motivated by CA models of self-reproduction, Tomita et al suggest that fixed, two-dimensional lattices – used as static backdrops to most conventional models – are unnecessarily restrictive for describing self-reproductive processes. They cite, as an example, the inability of conventional CA to describe biological processes (such as embryonic development) that must unfold in a finite closed space; once the underlying space of the CA is defined at the start, however large (and sometimes deliberately assumed *infinite*), its size remains the same throughout the development. This not only makes it hard to model the typically growing need that developing organisms have for *space*, but makes it impractical even to provide some room for avoiding overlaps between the original and daughter patterns [67].

Motivated by these, and other issues related to computation, Tomita et al. [67,69] introduce GRA, which is a form of graph grammar [20]. At first glance, GRA appear superficially similar to SDCA, at least in the sense that they both dynamically couple site values with topology. However, the transition rules are very different, and – in GRA's case – two properties hold that are not true for SDCA systems: (1) *all sites have exactly three neighbors at all times* (which is the minimum number of neighbors that yield nontrivial graphs [67]), and (2) *multiple links*



Structurally Dynamic Cellular Automata, Figure 17
Graphical representations of the actions of the GRA rules defined in Eq. (45). (Reproduced from [69] with permission)

are allowed to exist between any two sites. The authors claim that the 3-neighbor restriction not only does *not* constrain the space of emergent geometries (an observation that is echoed by Wolfram [75]; see Subsect. "Network Automata" below) but has the added benefit of allowing the rules to be expressed in a regular form: each rule is defined by a rule *name* and, at most, six *symbols* for its argument:

$$
\begin{aligned}
&(\sigma \text{ rules}):\\
&\quad\Big\{ \textit{transition}\,(x,a,b,c) \;\to\; (u,a,b,c),\\
&(\text{site rules}):\\
&\quad\begin{cases} \textit{division}\,(x,a,b,c) & \to\; (u,v,w,a,b,c),\\ \textit{fusion}\,(x,y,z,a,b,c) & \to\; (u,a,b,c), \end{cases}\\
&(\text{link rules}):\\
&\quad\begin{cases} \textit{commutation}\,(x,y,a,b,c,d) & \to\; (x,y,a,b,c,d),\\ \textit{annihilation}\,(x,y,a,b,c,d) & \to\; (a,b,c,d), \end{cases}
\end{aligned}
\tag{45}
$$

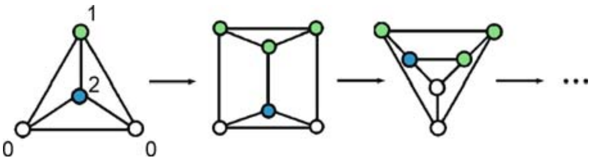
where x, y and z denote the σ values of the center sites *before* undergoing a structural change; u, v and w denote the σ values of the center sites *after* the structural change; and a, b, c, and d denote the states of the *neighboring sites*. The ordering is unimportant, so long as a given string can be obtained from another by cyclic permutation, otherwise the strings are different; i. e., (a,b,c) is both topologically and functionally equivalent to (b,c,a), but (c,b,a) is different. The action of σ, value, and links is graphically illustrated in Figure 17.

By convention, the GRA algorithm is applied in two steps: (1) site rules (*transition*, *division* and *fusion*) are executed first, and at all subsequent even time steps, followed by (2) link rules (*commutation* and *annihilation*), executed at odd steps.

In the event that multiple rules are simultaneously applicable – such as might happen, for example, if the rules include more than one division, or fusion, for the same lefthandside argument in their expressions (in Eq. (45)) – the order in which the rules are applied is determined by an a priori priority ranking. Also, since applying either *commutation* or *annihilation* rules to adjacent links yields inconsistency, whenever a local context arises in which this might happen, the application of these rules is temporarily suppressed. (This is done by sweeping through the link set *twice*: on the first pass, a temporary flag is set for each link that satisfies a rule condition; on the second pass, the link rule is applied *if and only if* the four neighboring links did not raise flags during the first pass.)

Figure 18 shows the first few steps in applying one *division* and two *commutation* rules to a simple initial graph. (Kohji Tomita provides several movies of GRA evolutions on his website: http://staff.aist.go.jp/k.tomita/ga/)

Tomita, Kurokawa, and Murata [67,68,69,70,71] report a variety of emergent behaviors, including (1) *arbitrary resolution* (because GRA rules effectively allow an arbitrary number of sites to "grow" out of any initial structure, these systems define their own "boundary conditions" and graphs with arbitrary resolution are possible); (2) *repetitive structures*, in which some geometrical subset of an initial graph is reproduced, indefinitely, and continu-



Structurally Dynamic Cellular Automata, Figure 18
Sample GRA evolution starting from the graph on the left. The rules are (see Eq. (45)): (1) *division*$(1,0,0,2) \to (1,1,1,0,0,2)$, (2) *commutation*$(1,2) \to (1,2)$, and (3) *commutation*$(0,0) \to (0,0)$. (Reproduced from [69] with permission)

ously grafted onto the original structure; and (3) *self-replication*, in which both site-value and structure is replicated after N steps. In [70], Tomita et al. describe how genetic algorithms [45] may be used for automating the search for self-replicating patterns.

In [67], Tomita et al. also present the design of a self-reproducing *Turing Machine*. *Turing machines* are abstract symbol manipulating devices that mimic the basic operations of a computer. Formally, they consist of a "tape" (of indefinite length, to record data), a "head" (that reads/writes symbols on the tape, and that can move left or right), and "state transition rules" (that tell the head which new symbols to write given the current state of the tape). The tape is analogous to "memory" in a modern computer; the head is analogous to the microprocessor. A Turing machine is called "universal" if it can simulate any other Turing machine.

Tomita et al.'s [67] *Turing machine* is modeled as a ladder structure: the upper sites constitute the "tape" mechanism; the lower sites form the "tape head" that reads the tape; both ends of the ladder are single sites that define "end of tape"; and the two ends are joined to form a loop. Although the tape is initially finite, the ladder can grow to arbitrary length, as required, by using appropriate GRA rules. Tomita et al. [67] self-replicating Turing GRA consists of 20 states and 257 (2-symbol) rules. They also introduce a design for a universal Turing machine [69] that consists of 30 states and 955 rules for reproduction, and 23 states and 745 rules for computation. While self-reproducing universal Turing machines can be described using conventional CA, their expression using GRA rules are considerably more compact.

Dynamic Graphs as Models of Self-Reconfigurable Robots

In the context of looking for self-reconfiguration algorithms that may be used to manufacture modular robots for industry, Saidani [63,64] has recently introduced a *dynamic graph calculus* that includes rules similar to those that define SDCA; but which depend only on the *topology* of (but not the σ-values living on) the lattice. Saidani and Piel [65] have also introduced an interactive programming environment for studying dynamic graph simulations called *Dynagraph*, and implemented in Smalltalk.

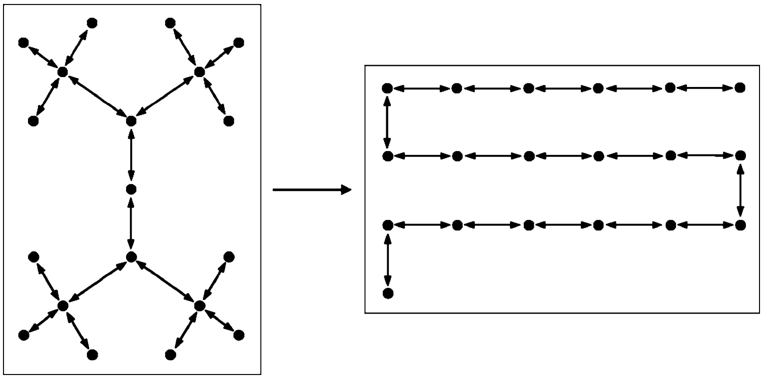
There are two basic approaches to designing modular robots: (1) to develop a set of elementary generic modules that can be rapidly assembled by humans to form robots that solve a specific problem, and (2) to design a set of (otherwise identical) primitive components that can adaptively reconfigure themselves. Focusing on the latter approach, Saidani [64] formally reinterprets modular "robots" to mean modular *networks*; and proceeds to model adaptive robotic self-reconfigurations as a class of recursive graph dynamical systems. In contrast to other related dynamic graph models [18,25], the "modules" (or subgraphs) of Saidani's model use local knowledge of their neighborhood topology to collectively evolve to some goal configuration. Although the dynamics transforms the global state, the evolution remains strictly decentralized, and individual modules do not know the (desired) final state.

Apart from restricting the dynamics to topology *alone* (indeed, none of the sites harbor information states of any kind), Saidani [63,64,65] further assumes that (1) connections between sites are *directional* (both *to*- and *from*-links may coexist between the same two modular components); (2) "active" sites reconfigure their local neighborhood by accepting, keeping, or removing their adjacent links according to rules that are functions of their current topology (defined as a given sites' current local neighborhood and the current neighborhood of its neighbors: a site only knows about its own *in*- and *out*-degree, which can obviously be computed from its local topology, and the *in*- and *out*-degrees of its nearest neighbors); (3) a site controls its *outgoing* links (and can connect or disconnect any outgoing links), but cannot sever incoming connections; (4) sites must maintain at least one link throughout an evolution (so that the graph remains connected); and (5) all sites are equipped with the same set of rules.

As in conventional CA and the basic SDCA model, the "reconfiguration" proceeds synchronously throughout the graph. The decision process includes an innate stochastic element: in the event that there is a rule that specifies that a site is to establish a link to a neighbor of one of its neighbors, but all neighboring sites have the same degree (which is the only dynamical discriminant), the neighbor with which a new link will be forged is selected at random.

As a concrete example, Saidani [64] presents a *tree-to-chain* algorithm that evolves an initial "tree" graph to a linear chain of linked sites (see Fig. 19). While we do not reproduce the full algorithm here, it is essentially a case-driven list of rules of the form *if condition* C_1 *(and condition* $C_2, \ldots$ *and condition* C_n*) then connect (or disconnect) site i to (from) the* n*th neighbor of i's neighbor*, ν. For example, an explicit "rule" might be: *if* $1 \le \deg^-(i) \le 2$ *and* $\deg^+(i) = 1$ *and* $|\tau(i)| = 2$ *then link i to a neighboring site j that has* $\deg^-(j) = 0$, where $\deg^-(i)$ and $\deg^+(i)$ are the *in*- and *out*-degrees of site *i*, and $\tau(i)$ is the total number of sites to which *i* is currently linked (with either incoming or outgoing links).

Conceptually, the details of Saidani's rules are less important than what the unfolding process represents as



Structurally Dynamic Cellular Automata, Figure 19
Schematic illustration of a *tree* topology reconfiguring itself into a linear *chain* using a set of case-based *"if–then"* topology rules defined in [64]; see text for details

a whole. An initial graph – which we recall is to be viewed as a distillation of a "modular robot" – is transformed, by the individual sites (or parts of the robot), into another *desired* structure; i. e., the graph is entirely *self-reconfigured*. Though the broader reverse-engineering problem (which includes asking such fundamental questions as *"How can a desired final state be mapped onto a specific cased-based list of graphical rules?"*) remains, as yet, unanswered, and the *Dynagraph* work environment [65] is currently limited to experimenting only with graphs that have less than 30 sites, the basic model already represents a viable new approach to using dynamic graphs to describe self-reconfigurable robots; and is potentially more far-reaching as a general model of topologically-reconfigurable dynamical systems.

SDCA as Models of Fundamental Physics

Pregeometric Theories of Emergent Space-Time

Although SDCA are a natural formal extension of conventional CA – and serve as general-purpose modeling tools – their conception was originally motivated by fundamental physics; specifically, by a search for models of self-organized emergent discrete space-time [42]. *"Space acts on matter, telling it how to move; … matter reacts back on space, telling it how to curve"*, which is the central lesson of Einstein's *geometrodynamics*, as explained by Misner, Thorne and Wheeler in their classic text on Gravitation [44]. Wheeler [73] has been a particularly eloquent spokesman for the need to search for what he calls a *pregeometry*, or a set of basic elements out of which what we normally think of geometry is built, but which are themselves devoid of a specific dimensionality: *"Space-time … often considered to be the ultimate continuum of physics, evidences nowhere more clearly than at big bang and at collapse that it cannot be a continuum. Obliterated in those events is not only matter, but the space and time that envelope that matter … we are led to ask out of what 'pregeometry' the geometry of space and spacetime are built"*. Wheeler has also proposed the idea that particles be viewed as geometric disturbances of spacetime, called *geometrodynamic excitons*.

A priori, SDCA appear tailor-made for describing pregeometric theories of space-time. Since in SDCA, lattice and local σ-values are explicitly coupled, and geometry and value configurations are treated on an approximately equal footing, SDCA is certainly at least formally consistent with Einstein's geometrodynamic credo. The structure is altered locally as a function of individual site neighborhood value-states and geometries, while local site-connectivity supports the site-value evolution in exactly the same way as in conventional CA models defined on random lattices. The microphysical view of physics that emerges from this construction is one in which a fundamentally discrete pregeometry continually evolves in time as an amorphous structure but with a globally well-defined dimensionality. Particles are constructs of that amorphous structure and can be viewed as locally persistent substructures – i. e. geometrical or topological solitons – with dimensions that differ from the surrounding value. Just as "value structure" solitons are ubiquitous in conventional CA models [32,74], "link structure" solitons might emerge in SDCA; physical particles would, in such a scheme, be viewed as geometrodynamic disturbances propagating within a dynamic lattice.

Of course, speculation regarding the ultimate constituents of matter and space-time date back at least as far as 500 BC when the philosopher Democritus mused on the idea that matter is made of indivisible units separated by void. Since then there have been countless attempts, with varying degrees of success, to fashion an entirely discrete theory of nature. We limit our discussion to a short survey of some recent work that centers on ideas that are either direct outgrowths of, or are otherwise conceptually related to, SDCA models. (A short history of pregeometric theories appears in chapter twelve of Ilachinski [32]).

One of the earliest proponents of pregeometry is Zuse [76], who speculated on what it would take for a CA-like universe to sustain "digital particles" on a cellular lattice. He focused on two main problems: (1) *How does the universe's observed isotropy arise from a CA's (Euclidean, hexagonal, etc.) anisotropy?*, and (2) *What is the information content of a physical particle?* As an answer to the first question, Zuse suggests …

"… *variable and growing automata. Irregularities of the grid structure are a function of moving patterns, which is represented by digital particles. Now, not only certain values are assigned to the single crosspoints of the grid in the concept of the cellular automaton which are interrelated and sequencing each other, but also the irregularities of the grid are itself functions of these values of the just existing interlinking network. One can imagine rather easily that in such a way the interdependence of mass, energy, and curvature of space may logically result from the behavior of the grid structure.*"

Jourjine [33] generalizes Euclidean lattice field theory on a d-dimensional lattice to a cell complex. Using homology theory to replace points by cells of various dimensions and fields by functions on cells, he develops a formalism that treats space-time as a dynamical variable and describes the change in the dimension of space-time as a phase transition.

Kaplunovsky and Weinstein [34] develop a field-theoretic formalism that treats the topology and dimension of the spacetime continuum as dynamically generated variables. Dimensionality is introduced out of the characteristic behavior of the energy spectrum of a system of a large number of coupled oscillators.

Dadic and Pisk [13] introduce a self-generating discrete-space model that is based on the local quantum-mechanics of graphs. Just as in SDCA, Dadic and Pisk's spatial structure is discrete but not static; it is fundamentally amorphous and evolves in time. Though the metric is essentially the same one used to define SDCA (i. e., D_{effec}), it is generalized to unlabeled graphs by referring to the topological description of the node positions rather than their arbitrary labels. Though their "graph dynamics" differs from what is used by SDCA (and uses a symmetrized Fock space that is local in terms of their graph metric, where "Fock space" is a Hilbert space used to describe quantum states with a variable, or unspecified, number of particles, and is made from the direct sum of tensor products of single-particle; or, in this case, single-*graph*, Hilbert spaces) it shares two important properties with SDCA: (1) interactions depend only on the local properties of the graph, and (2) interactions induce only minimal changes to the local metric function. An important consequence of their theory is that the dimension of a graph is a scale dependent quantity that is generated by the dynamics.

Combinatorial Space-Time Hillman [27] introduces a *combinatorial space-time*, which he defines as a class of dynamical systems in which finite pieces of spacetime contain finite amounts of information. Spacetime is modeled as a combinatorial object, constructed by dynamically coupling copies of finitely many types of certain allowed neighborhoods. There is no *a priori* metric, and no concept of continuity, which is expected to emerge on the macroscale.

The construction (and evolution) of spaces proceeds in three steps: (1) define a set X of combinatorial n-dimensional spaces (examples are conventional CA graphs, graphs with directional links, or some other kind of embedded symmetry); (2) define a set of local, invertible *primitive* maps $T\colon X \leftrightarrow Y$ between pairs of space sets, such that the maps do not all commute with one another (for example, a simple renaming of the sites or links gives an invertible, local map); (3) generate an arbitrary set of local invertible graph transformations by composing primitive maps with one another. Since the primitive maps are deliberately chosen so that they do not all commute, the act of composition yields infinitely many nontrivial transformations. The *orbits* $\{T^z(x) \mid z \in Z\}$ (for each space x in X) are $(n+1)$-dimensional combinatorial spacetimes; which include reversible CA and SDCA-like networks in which geometry evolves locally over time. Formally, Hillman uses matrices of nonnegative integers, directed graphs, and symmetric tensors to describe these systems, so that local equivalences between space sets are generated by simple matrix transformations. Concrete examples of dynamic combinatorial space-time graphs are given in [27].

Structurally Dynamic Disordered Cellular Networks As an explicit example of how dynamic graphs can be used to model pregeometry, consider *structurally dynamic disordered cellular networks* (abbreviated, SDDCN), recently introduced by Nowotny and Requardt [53,54,55,58,59,60]. SDDCN are a class of models closely related to SDCA but developed explicitly to describe a discrete, dynamic spacetime fundamental physics. The main difference between the two models is that whereas link connections in SDCA are strictly local, SDDCN are capable of generating both local and *translocal* links.

In contrast to more mainstream high-energy theories of fundamental physics (which are dominated by string theory and/or loop quantum gravity, both of which assume a certain level discretization at the Planck scale, but assume that a discrete space-time emerges from an underlying *continuum* physics), SDDCN takes a *bottom-up* approach. SDDCN assumes that there is underlying dynamic, discrete and highly erratic network substratum that consists of (on a given scale) irreducible mutually interacting agents exchanging information via primordial channels (links). The known continuum structures are expected to emerge on a macroscopic (or, mesoscopic) scale, via a sequence of coarse graining and/or renormalization steps.

Like SDCA, SDDCN are defined on arbitrary graphs, G, initially defined by a specified set of sites and links. Both sites and links are allowed to take on values. Site values, σ_i (which represent a primitive "charge"), are taken from some discrete set, $q \cdot \mathcal{Z}$, where q is a discrete *quantum of information*; link states assume the values $J_{ij} \in \{-1, 0, +1\}$, and represent an elementary coupling. The J_{ij} are equivalent to SDCA's l_{ij}, but take on *three* values rather than two. Heuristically, J_{ij} represent directed edges pointing either from site i to j (if $J_{ij} = 1$), or from j to i (if $J_{ij} = -1$); or, in the case of $J_{ij} = 0$, the absence of a link. At each time step (representing an elementary quantum of time), an elementary quantum q is transported along each existing directed link in the indicated direction. As for SDCA, SDDCN dynamically couples site values to links.

Nowotny and Requardt [53] introduce two network models: one in which connected sites that have very different internal states typically lead to large local fluctuations (= SDDCN$_1$), and another in which sites with similar internal states are connected (= SDDCN$_2$):

SDDCN$_1$

$$\leftrightarrow \begin{cases} \sigma_i^{t+1} = \sigma_i^t + \sum_j J_{ji}^t , & \\ J_{ij}^{t+1} = \mathrm{sign}\left(\Delta\sigma_{ij}\right) & \text{for} \begin{cases} \left|\Delta\sigma_{ij}\right| \geq \lambda_2 , \text{ or} \\ \left|\Delta\sigma_{ij}\right| \geq \lambda_1 \wedge J_{ij}^t \neq 0 , \end{cases} \\ J_{ij}^{t+1} = 0 , & \text{otherwise,} \end{cases}$$

SDDCN$_2$

$$\leftrightarrow \begin{cases} \sigma_i^{t+1} = \sigma_i^t + \sum_j J_{ji}^t , & \\ J_{ij}^{t+1} = \mathrm{sign}\left(\Delta\sigma_{ij}\right) & \text{for} \begin{cases} 0 < \left|\Delta\sigma_{ij}\right| < \lambda_1 , \text{ or} \\ 0 < \left|\Delta\sigma_{ij}\right| < \lambda_2 \wedge J_{ij}^t \neq 0 , \end{cases} \\ J_{ij}^{t+1} = J_{ij}^t , & \text{for } \Delta\sigma_{ij} = 0 , \\ J_{ij}^{t+1} = 0 , & \text{otherwise ,} \end{cases} \tag{46}$$

where $\Delta\sigma_{ij} = \sigma_i^t - \sigma_j^t$, and $\lambda_2 \geq \lambda_1 \geq 0$. Since SDDCN is intended to model pregeometric dynamics, Nowotny and Requardt [53] caution that the t parameter that appears in these equations must not to be confused with the "true time" that (they expect) emerges on coarser scales. In keeping with its physics-based motivation, SDDCN's dynamical laws depend only on the relative differences in site values, not on their absolute values. Indeed, charge is nowhere either created or destroyed, so that SDDCN conserves global "charge": $\sum_i \sigma_i^t =$ constant, where the arbitrary constant can be set to *zero*.

Both models start out initially on a simplex graph with $N \sim 200$ nodes, so that the maximum number of possible links is $N(N-1)/2$. The initial σ-seed consists of a uniform random distribution of values scattered over the interval $\{-k, -k+1, \ldots, k-1, k\}$, where $k \sim 100$. The initial values for link states, $J_{ij}^{t=0}$, are selected from $\{-1, 1\}$ with equal probability; i. e., the initial state is a maximally entangled nucleus of nodes and links. Nowotny and Requardt [55] state that "*… in a sense, this is a scenario which tries to imitate the big bang scenario. The hope is, that from this nucleus some large-scale patterns may ultimately emerge for large clock-time*". For most properties (other than the $\langle\sigma\rangle_t$ and $\sum_i |\sigma_i^{t+1} - \sigma_i^t|$, which are both equal to *zero* by construction), the average over the width of the initial vertex state distribution, taken over λ_1 and λ_2, specific realizations of initial conditions, and time, depend linearly on network size.

We summarize Nowotny's and Requardt's [53,55] findings, culled from extensive numerical experiments: (1) the appearance of very short limit cycles in SDDCN$_1$ (period 6 and multiples of 6, with the longest having period 36 on a network of size $N = 800$), (2) Much longer limit cycles and transients in SDDCN$_2$, both of which appear to grow approximately exponentially, (3) structurally, SDDCN$_1$ evolve from almost fully connected simplex networks to more sparse connectivities with increasing $\lambda_{1/2}$; there is a regime in which few vertices with very high degree coexist with many vertices with a low degree; for large (around $\lambda_1 \approx 60$; for large $\lambda_{1/2}$, the graph eventually breaks apart and all nodes become isolated; (4) for SDDCN$_2$, nodes typically have zero degree small $\lambda_{1/2}$, and links become increasingly dense as $\lambda_{1/2}$ increase; the degree distribution is generally broad and remains so for large $\lambda_{1/2}$ (the authors also note observing multiple local maxima of the distributions in a wide range of $\lambda_{1/2}$ values); (5) for SDDCN$_1$, there is an abrupt phase-transition in the temporal fluctuations of vertex degrees (defined as $\deg_i(t+1) - \deg_i(t)$) from a state in which there are essentially no fluctuations ("frozen network") to one with strong fluctuations ("liquid network"); (6) the distribution of site values is strongly bimodal for $62 \leq \lambda_1 \leq 85$ for SDDCN_1 (while SDDCN_1 distributions are not bimodal, the width of the site value distributions for different values of λ_1 appears modulated.

From a fundamental physics perspective, the most interesting class of behaviors of SDDCN involves emergent dimensionality. Nowotny and Requardt [55] argue that since the continuum is a self-organized dynamic structure that emerges in the limit of large N and t, the most useful measure of "dimension" cannot be purely local (as in the case of *effective dimensionality*, D_{effec}, used for describing SDCA systems). Rather, it must be an intrinsically *global property*; one that is independent of any arbitrary embedding dimension, and one that can take on relatively stable values in the *whole* (to characterize effective system-wide characteristics), while simultaneously being relatively impervious to otherwise rapidly changing structural changing taking place on the microscale. Toward this end, Nowotny and Requardt [53] define the upper (and lower) scaling dimensions, $D_S^U(i)$ (and $D_S^L(i)$), with respect to site i:

$$\begin{aligned} D_S^U(i) &= \limsup_{r\to\infty} \frac{\ln \beta(i,r)}{\ln r}, \\ D_S^L(i) &= \liminf_{r\to\infty} \frac{\ln \beta(i,r)}{\ln r}, \end{aligned} \tag{47}$$

and the upper (and lower) connectivity dimensions, $D_C^U(i)$ (and $D_C^L(i)$), with respect to site i:

$$\begin{aligned} D_C^U(i) &= \limsup_{r\to\infty} \frac{\ln \partial\beta(i,r)}{\ln r}, \\ D_C^L(i) &= \liminf_{r\to\infty} \frac{\ln \partial\beta(i,r)}{\ln r}, \end{aligned} \tag{48}$$

where $\beta(i,r) = \#\,\text{sites} j \mid D_{ij} \leq r$, and $\partial\beta(i,r)$ is the number of sites on the *surface* of the r-sphere. When the upper and lower limits coincide, we have the *scaling dimension* $(= D_S)$ and the *connectivity dimension* $(= D_C)$, respectively. D_S is related to well known dimensional concepts in fractal geometry; D_C is a more physical measure that describes how the graph is connected, and thus how sites may potentially influence one another [55]. Preliminary research [53] suggests that under certain conditions, behavior resembling a structural phase transition to states with stable internal (and/or connectivity) dimensions is possible.

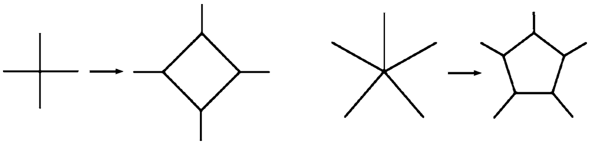
Network Automata

Stephen Wolfram devotes chapter nine of his Opus – *A new kind of science* (abbreviated, NKS) [75] – to applying CA to fundamental physics; and speculates on ways in which space may be described using a dynamic network. The central, overarching theme of NKS is that "simple" programs often suffice to capture complex behaviors.

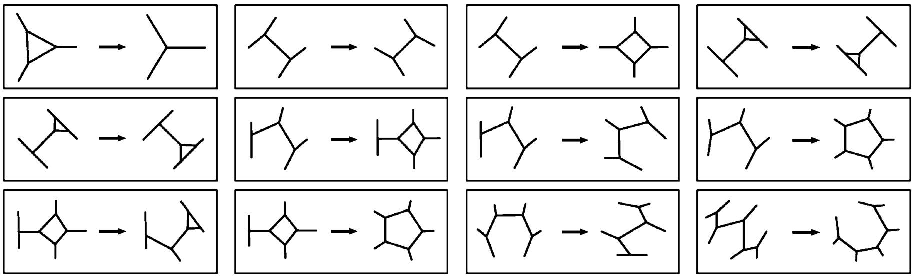
The bold claim made in chapter nine of NKS is that, on an even more fundamental level, what underlies *all the laws of physics*, as we currently understand them, is a simple CA-like program, from which, ultimately, all the phenomenologically observed complexity in the universe naturally emerges. As for the specific forms such a "program" may take, Wolfram's intellectual point of departure echoes that of other proponents of a discrete dynamic pregeometric theory:

"... cellular automata ... cells are always arranged in a rigid array in space. I strongly suspect that in the underlying rule for our universe there will be no such built-in structure. Rather ... my guess is that at the lowest level there will just be certain patterns of connectivity that tend to exist, and that space as we know it will then emerge from these patterns as a kind of large-scale limit".

Wolfram introduces his *network automata* (abbreviated, NA) with these basic assumptions (see additional notes in NKS [75] on the evolution of networks: pp. 1037–1040): (1) features of our universe emerge solely from properties of space, (2) the underlying model (and/or "rules") must contain only a minimal underlying geometric structure, (3) the individual sites of emergent graphs must not be assigned any intrinsic position, (4) sites are limited to possessing purely *topological* information (that defines the set of sites to which a given site is connected), (5) incoming and outgoing connections need not be distinguished, and (6) all sites have exactly the same total number of links to other sites (which is assumed equal to *three*). This last assumption – which is essentially the same one made by Nowotny and Requardt [53] as the basis of their SDDCN model; see Subsect. "Structurally Dynamic Disordered Cellular Networks" above – does not lead to any loss of generality. With two connections, only very trivial graphs are possible; and it is easy to show that any site with more than three links can always be redefined, locally, as a collection of sites with exactly three links each (see Fig. 20).



Structurally Dynamic Cellular Automata, Figure 20
Illustration of how sites that have more than three links can always be redefined as a set of sites with exactly three links each



Structurally Dynamic Cellular Automata, Figure 21
Examples of planarity-preserving network substitution rules. (Reproduced from [75] with permission)

Wolfram [75] gives several concrete examples of evolving graphs (as models of pregeometry), the dynamics of which are prescribed by a set of *substitution* rules rules; i. e., explicit lists of the topological configurations (of sites and links) that are used to replace (at time $t+1$) specific local configurations (as they appear at time t). However, in contrast to SDCA rules, Wolfram's substitution rules are strictly *topological*; no site-value information is used. Also, the number of sites in the graph can change as the graph evolves; where, in SDCA, the number remains constant.

Figure 21 shows examples of rules in which specific clusters of sites are replaced with other clusters of sites. While the rules shown in the figure share the property that they all preserve planarity, there is no particular reason for imposing such a restriction; in fact, rules that generate non-planarity are just as easy to define. Wolfram speculates (pp. 526–530 in [75]) that "particle states" may be defined as mobile *non*-planar subgraphs that persist on an otherwise planar, but *randomly* fluctuating topology. Reversible versions of these rules may also be constructed, by associating a "backward" version with each "forward" transformation.

Some care must be taken while both *defining* and *applying* these rules consistently. For example, if a cluster of sites contains a certain number of links at t, one is not permitted to define a rule that replaces that cluster with another one that has a different number of connections. Another restriction is that rules must be independent of orientation; that is, if a candidate rule requires identifying the specific links (of, say, an otherwise topologically symmetric n-link local subgraph) before activating a desired substitution, that rule is likewise forbidden. However, even with these restrictions, a large number of rules are still possible. For example, 419 distinct rules may be defined for clusters with no more than five sites.

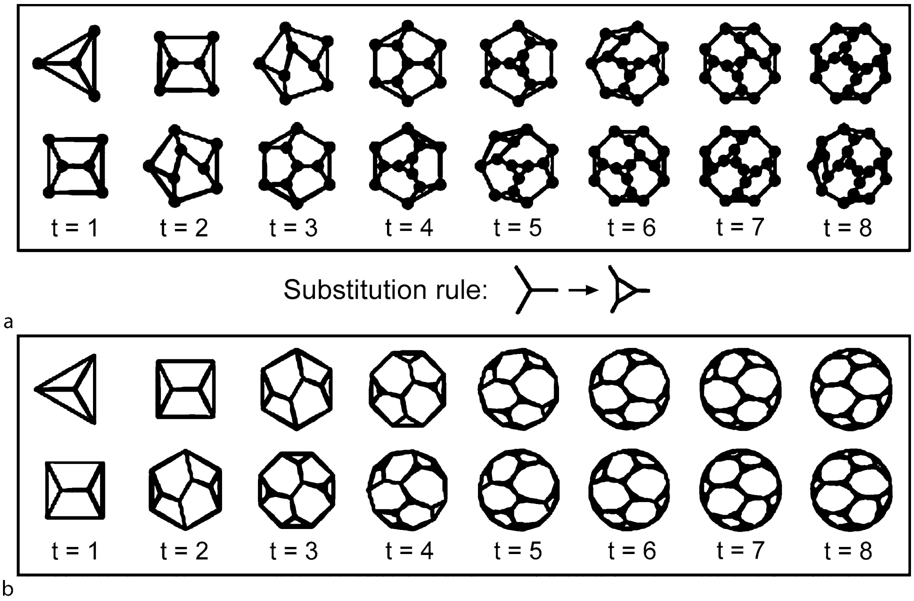
In applying network rules, one cannot simply simultaneously replace all pertinent subgraphs with their replacements, since, in general, two or more subgraphs with the same topology may overlap somewhere within the network. Since there is no *priori*, or universally consistent, way of ordering the subgraphs, *meta*-rules must be imposed to eliminate any possible ambiguities. For example, one method (*m1*) is to restrict replacements to a single subgraph per time step, selecting the subgraph whose replacement entails the minimal change to all recently updated sites. Another method (*m2*) is to allow all possible nonoverlapping replacements, while ignoring those that overlap. Wolfram reports that, although the second method obviously produces larger graphs in fewer steps, the two methods generally produce qualitatively similar structures.

Figure 22 traces the first few steps in the evolution of a simple graph under the action of a single substitution rule (defined at the center of the figure). Figures 22a and 22b show the results of applying this rule using methods *m1* and *m2*, respectively. In each case, the top row shows the form of the network before the substitution takes place at that step, and the bottom row shows the network that results from the substitution. The subgraph (or subgraphs, in Fig. 22a) involved in the replacement is highlighted at both top and bottom.

Wolfram also suggests that analogs of *mobile automata* [43] can be defined for evolving networks. By tagging a site i, say, with a "charge", $\sigma_i \equiv 1$, substitution rules may be defined to replace clusters of sites around the charged site. The effect is that the charge itself appears to move, as its effective (relative) position within the network changes as the geometric dynamics unfolds. (However, Wolfram also notes – on page 1040 in [75] – that *"despite looking at several hundred cases I have not been able to find network mobile automata with especially complicated behavior"*).

Future Directions and Speculations

Although SDCA were first introduced over two decades ago [28], much of their behavior remains unexplored. Of course, this is due largely to the difficulty of studying dy-



Structurally Dynamic Cellular Automata, Figure 22
Examples of network evolutions using the substitution rule shown at center. See text for explanation. (Reproduced from [75] with permission)

namical systems that harbor an a priori vastly larger coupled value-geometry space than the "merely" spatially-confined behavioral space of conventional CA. Only relatively recently have desktop computers become sufficiently powerful, and visualization programs adept enough at rendering multidimensional graphs [12], to make a serious study of SDCA behaviors possible. For example, the general-purpose math programs *Mathematica* (http://www.wri.com) and *Maple* (http://www.maplesoft.com) both provide powerful built-in graph-rendering algorithms to help visualize complex graphs. Standalone public-domain packages are also available; for example, AGNA [2], NetDraw [11], and Pajek [52]. In this final section, we list several open questions and briefly speculate on possible future directions.

Because of the relative paucity of studies dedicated purely to exploring the space of emergent structures (such as Wolfram's [74] pioneering studies of conventional CA), many (even very fundamental) questions remain open: *What kinds of geometries can arise?, Which subspace of the space of all possible graphs corresponds to those that are actually attainable using SDCA (and SDCA-like) rules?, What are the conditions for which certain geometries do, and do not, form?, What combinations of* σ*- and* ℓ*-rules give rise to specific kinds of graphs?*

Other open problems include: (1) determining whether the (provisionally defined) set of class-4 rules, for which effective dimension appears to remain constant, is *genuine*, rather than being either a long-term transient or an unintentional artifact of imposed run-time constraints; and, if this class is "real", we obviously need to ask, *How large is it?*, and *Under what conditions does it arise?*; (2) developing SDCA as formal mathematical models, perhaps as members of a broader class of graph grammars [20,35]; and (3) finding purely geometric analogs of the solitons known to exist in conventional CA models [32,74].

This article has introduced several generalizations of the basic SDCA model, including *memory* effects (Subsect. "SDCA With Memory"), *reversibility* (Subsect. "Reversible SDCA"), *probabilistic transitions* (Subsect. "Probabilistic SDCA"), and a class of SDCA-like dynamical systems that evolve according to rules that depend only on *topology* (Subsects. "Dynamic Graphs as Models of Self-Reconfigurable Robots" and "Network Automata"). However, other possibilities abound: (1) σ site-variables may take on a larger range of values, $\sigma \in \{0, 1, \ldots, k-1\}$; (2) link variables, ℓ_{ij}, may similarly take on a larger range of values, $\ell_{ij} \in \{0, \pm 1, \pm 2, \ldots, \pm m\}$ (where, say, ± determines "directionality", and absolute value, $|\ell_{ij}|$, represents either channel capacity for information flow or some other innate property); and (3) both sites and links may take on richer, and more explicitly "active", roles of *agent-actors* [17].

Apart from these formal extensions, some obvious future applications include modeling communication and social network dynamics, studying the dynamics of plasticity in artificial neural networks, designing adaptive self-reconfiguring parallel-computer networks (as well as "amorphous" computer chips), studying behaviors of gene-regulatory networks, and providing the conceptual core for fundamental pregeometric physical theories of discrete, emergent space-times.

Bibliography

Primary Literature

1. Adamatzky A (1995) Identification of Cellular Automata. Taylor and Francis, London
2. Applied Graph & Network Analysis software. http://benta.addr.com/agna/. Accessed 14 Oct 2008
3. Albert J, Culik II K (1987) A simple universal cellular automaton and its one-way and totalistic version. Complex Syst 1:1–16
4. Ali SM, Zimmer RM (1995) Games of Proto-Life in Masked Cellular Automata. Complexity International, vol 2. http://www.complexity.org.au
5. Ali SM, Zimmer RM (2000) A Formal Framework for Emergent Panpsychism. In: Tucson 2000: Consciousness Research Abstracts. http://www.consciousness.arizona.edu/tucson2000/. Accessed 14 Oct 2008
6. Alonso-Sanz R, Martin M (2006) A structurally dynamic cellular automaton with memory in the hexagonal tessellation. In: El Yacoubi S, Chopard B, Bandini S (eds) Lecture Notes in Computer Science, vol 4173. Springer, New York, pp 30–40
7. Alonso-Sanz R (2006) The Beehive Cellular Automaton with Memory. J Cell Autom 1(3):195–211
8. Alonso-Sanz R (2007) A structurally dynamic cellular automaton with memory. Chaos, Solitons, Fractals 32(4):1285–1304
9. Barabasi AL, Albert R (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97
10. Bollobas B (2002) Modern Graph Theory. Springer, New York
11. Borgatti SP (2002) NetDraw 1.0: Network visualization software. Analytic Technologies, Harvard
12. Chen C (2004) Graph Drawing Algorithms. In: Information Visualization. Springer, New York
13. Dadic I, Pisk K (1979) Dynamics of discrete-space structure. Int J Theor Phys 18:345–358
14. Doi H (1984) Graph theoretical analysis of cleavage pattern: graph developmental system and its application to cleavage pattern in ascidian egg. Dev Growth Differ 26(1):49–60
15. Durrett R (2006) Random Graph Dynamics. Cambridge University Press, New York
16. Erdos P, Renyi A (1960) On the evolution of random graphs. Publ Math Inst Hung Acad Scie 5:11–61
17. Ferber J (1999) Multi-Agent Systems: An Introduction to Distributed Artificial Intelligence. Addison-Wesley, New York
18. Ferreira A (2002) On models and algorithms for dynamic communication networks: the case for evolving graphs. In: 4th Recontres Francophones sur les Aspects Algorithmiques des Télécommunications (ALGOTEL 2002), Meze, France
19. Gerstner W, Kistler WM (2002) Spiking Neuron Models. Single Neurons, Populations, Plasticity. Cambridge University Press, New York
20. Grzegorz R (1997) Handbook of Graph Grammars and Computing by Graph Transformation. World Scientific, Singapore
21. Halpern P (1989) Sticks and stones: a guide to structurally dynamic cellular automata. Amer J Phys 57(5):405–408
22. Halpern P, Caltagirone G (1990) Behavior of topological cellular automata. Complex Syst 4:623–651
23. Halpern P (1996) Genetic algorithms on structurally dynamic lattices. In: Toffo T, Biafore M, Leao J (eds) PhysComp96. New England Complex Systems Institute, Cambridge, pp 135–136
24. Halpern P (2003) Evolutionary Algorithms on a Self-Organized, Dynamic Lattice. In: Bar-Yam Y, Minai A (eds) Unifying Themes in Complex Systems, vol 2. Proceedings of the Second International Conference on Complex Systems. Westview Press, Cambridge
25. Harary F, Gupta G (1997) Dynamic graph models. Math Comp Model 25(7):79–87
26. Hasslacher B, Meyer D (1998) Modeling dynamical geometry with lattice gas automata. Int J Mod Phys C 9:1597
27. Hillman D (1995) Combinatorial Spacetimes. Ph D Dissertation, University of Pittsburg
28. Ilachinski A (1986) Topological life-games I. Preprint. State University of New York at Stony Brook
29. Ilachinski A, Halpern P (1987) Structurally dynamic cellular automata. Preprint. State University of New York at Stony Brook
30. Ilachinski A, Halpern P (1987) Structurally dynamic cellular automata. Complex Syst 1(3):503–527
31. Ilachinski A (1988) Computer Explorations of Self Organization in Discrete Complex Systems. Diss Abstr Int B 49(12):5349
32. Ilachinski A (2001) Cellular Automata: A Discrete Universe. World Scientific, Singapore
33. Jourjine AN (1985) Dimensional phase transitions: coupling of matter to the cell complex. Phys Rev D 31:1443
34. Kaplunovsky V, Weinstein M (1985) Space-time: arena or illusion? Phys Rev D 31:1879–1898
35. Kniemeyer O, Buck-Sorlin GH, Kurth W (2004) A Graph Grammar Approach to Artificial Life. Artif Life 10(4):413–431
36. Krivovichev SV (2004) Crystal structures and cellular automata. Acta Crystallogr A 60(3):257–262
37. Lehmann KA, Kaufmann M (2005) Evolutionary algorithms for the self-organized evolution of networks. In: Proceedings of the 2005 Conference on Genetic and Evolutionary Computation, Washington DC. ACM Press, New York
38. Love P, Bruce M, Meyer D (2004) Lattice gas simulations of dynamical geometry in one dimension. Phil Trans Royal Soc A: Math Phys Eng Sci 362(1821):1667–1675
39. Majercik S (1994) Structurally dynamic cellular automata. Master's Thesis, Department of Computer Science, University of Southern Maine
40. Makowiec D (2004) Cellular Automata with Majority Rule on Evolving Network. In: Lecture Notes in Computer Science, vol 3305. Springer, Berlin, pp 141–150
41. Mendes RV (2004) Tools for network dynamics. Int J Bifurc Chaos 15(4):1185–1213
42. Meschini D, Lehto M, Piilonen J (2005) Geometry, pregeometry and beyond. Stud Hist Phil Mod Phys 36:435–464
43. Miramontes O, Solé R, Goodwin B (1993) Collective behavior of random-activated mobile cellular automata. Physica D 63:145–160
44. Misner CW, Thorne KS, Wheeler JA (1973) Gravitation. W.H. Freeman, New York
45. Mitchell M (1998) An Introduction to Genetic Algorithms. MIT Press, Boston

46. Moore EF (1962) Sequential Machines: Selected Papers. Addison-Wesley, New York
47. Muhlenbein H (1991) Parallel genetic algorithm, population dynamics and combinatorial optimization. In: Schaffer H (ed) Third International Conference on Genetic Algorithms. Morgan Kauffman, San Francisco
48. Murata S, Tomita K, Kurokawa H (2002) System generation by graph automata. In: Ueda K (ed) Proceedings of the 4th International Workshop on Emergent Synthesis (IWES '02), Kobe University, Japan, pp 47–52
49. Mustafa S (1999) The Concept of Poiesis and Its Application in a Heideggerian Critique of Computationally Emergent Artificiality. Ph D Thesis, Brunel University, London
50. Newman M, Barabasi A, Watts DJ (2006) The Structure and Dynamics of Networks. Princeton University Press, New Jersey
51. Nochella J (2006) Cellular automata on networks. Talk given at the Wolfram Science Conference (NKS2006), Washington DC, USA, 16–18 June
52. Nooy W, Mrvar A, Batagelj V (2005) Exploratory Social Network Analysis with Pajek. Cambridge University Press, New York
53. Nowotny T, Requardt M (1998) Dimension Theory of Graphs and Networks. J Phys A 31:2447–2463
54. Nowotny T, Requardt M (1999) Pregeometric Concepts on Graphs and Cellular Networks as Possible Models of Space-Time at the Planck-Scale. Chaos, Solitons, Fractals 10:469–486
55. Nowotny T, Requardt M (2006) Emergent Properties in Structurally Dynamic Disordered Cellular Networks. arXiv: cond-mat/0611427. Accessed 14 Oct 2008
56. O'Sullivan D (2001) Graph-cellular automata: a generalized discrete urban and regional model. Environ Plan B: Plan Des 28(5):687–705
57. Prusinkiewicz P, Lindenmayer (1990) The Algorithmic Beauty of Plants. Springer, New York
58. Requardt M (1998) Cellular Networks as Models for Planck-Scale Physics. J Phys A 31:7997–8021
59. Requardt M (2003) A geometric renormalisation group in discrete quantum space-time. J Math Phys 44:5588–5615
60. Requardt M (2003) Scale free small world networks and the structure of quantum space-time. arXiv.org:gr-qc/0308089
61. Rose H (1993) Topologische Zellulaere Automaten. Master's Thesis, Humboldt University of Berlin
62. Rose H, Hempel H, Schimansky-Geier L (1994) Stochastic dynamics of catalytic CO oxidation on Pt(100). Physica A 206:421–440
63. Saidani S (2003) Topodynamique de Graphe. Les Journées Graphes, Réseaux et Modélisation. ESPCI, Paris
64. Saidani S (2004) Self-reconfigurable robots topodynamic. In: IEEE International Conference on Robotics and Automation, vol 3. IEEE Press, New York, pp 2883–2887
65. Saidani S, Piel M (2004) DynaGraph: a Smalltalk Environment for Self-Reconfigurable Robots Simulation. European Smalltalk User Group Conference. http://www.esug.org/
66. Schliecker G (1998) Binary random cellular structures. Phys Rev E 57:R1219–R1222
67. Tomita K, Kurokawa H, Murata S (2002) Graph Automata: Natural Expression of Self-Reproduction. Physica D 171(4):197–210
68. Tomita K, Kurokawa H, Murata S (2005) Self-description for construction and execution in graph rewriting automata. In: Lecture Notes in Computer Science, vol 3630. Springer, Berlin, pp 705–715
69. Tomita K, Kurokawa H, Murata S (2006) Two-state graph-rewriting automata. NKS 2006 Conference, Washington, DC
70. Tomita K, Kurokawa H, Murata S (2006) Automatic generation of self-replicating patterns in graph automata. Int J Bifurc Chaos 16(4):1011–1018
71. Tomita K, Kurokawa H, Murata S (2006) Self-Description for Construction and Computation on Graph-Rewriting Automata. Artif Life 13(4):383–396
72. Weinert K, Mehnen J, Rudolph G (2002) Dynamic Neighborhood Structures in Parallel Evolution Strategies. Complex Syst 13(3):227–244
73. Wheeler JA (1982) The computer and the universe. Int J Theor Phys 21:557
74. Wolfram S (1984) Universality and complexity in cellular automata. Physica D 10:1–35
75. Wolfram S (2002) A New Kind of Science. Wolfram Media, Champaign, pp 508–545
76. Zuse K (1982) The Computing Universe. Int J Theor Phys 21:589–600

Books and Reviews

Battista G, Eades P, Tamassia R, Tollis IG (1999) Graph Drawing: Algorithms for the Visualization of Graphs. Prentice Hall, New Jersey

Bornholdt S, Schuster HG (eds) (2003) Handbook of Graphs and Networks. Wiley-VCH

Breiger R, Carley K, Pattison P (2003) Dynamical Social Network Modeling and Analysis. The National Academy Press, Washington DC

Dogogovtsev SN, Mendes JF (2003) Evolution of Networks. Oxford University Press, New York

Durrett R (2006) Random Graph Dynamics. Cambridge University Press, New York

Gross JL, Yellen J (eds) (2004) Handbook of Graph Theory. CRC Press, Boca Raton

Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy

Benjamin A. Brooks[1], James H. Foster[1], Jeffrey J. McGuire[2], Mark Behn[2]
[1] School of Ocean and Earth Science and Technology, University of Hawaii, Honolulu, USA
[2] Department of Geology and Geophysics, Woods Hole Oceanographic Institution, Woods Hole, USA

Article Outline

Glossary

Submarine landslide A gravitational mass failure feature on the seafloor.

Slow earthquake A discrete slip event that produces millimeter to meter-scale displacements identical to those produced during earthquakes but without the associated seismic shaking.

GPS The Global Positioning System consists of a constellation of at least 24 medium earth orbiting satellites transmitting two or more microwave frequencies for use in precise positioning.

Seafloor geodesy The application of geodetic methods (studies of the change in the shape of the earth's surface) applied to a submarine environment.

Definition of the Subject

The term 'submarine landslide' encompasses a multitude of gravitational mass failure features at areal scales from square meters to thousands of square kilometers. Here, we concentrate on the large end of that spectrum, namely the submarine landslides that, when they move either in contained slip events or catastrophically, can generate surface displacements equivalent to $> M6$ earthquakes and/or hazardous tsunami.

The term 'slow earthquake' describes a discrete slip event that produces millimeter to meter-scale displacements identical to those produced during earthquakes but without the associated seismic shaking. Slow earthquakes, primarily associated with tectonic fault zones, have been recognized and studied with increasing frequency in the past decade largely due to the decreasing cost and proliferation of Global Positioning System (GPS) geodetic networks capable of detecting the ground motion [1,2,3]. Recently, one such GPS network on the south flank of Kilauea volcano, has recorded multiple slow earthquakes on the subaerial portion of a large landslide system that occurs primarily in the submarine environment [4,5,6]. Because the bathymetric charts surrounding the Hawaiian islands are littered with the remnants of massive, catastrophically emplaced submarine landslides (Fig. 1) it is natural to wonder if a slow-slipping submarine landslide is a precursory stage of one that will ultimately fail catastrophically.

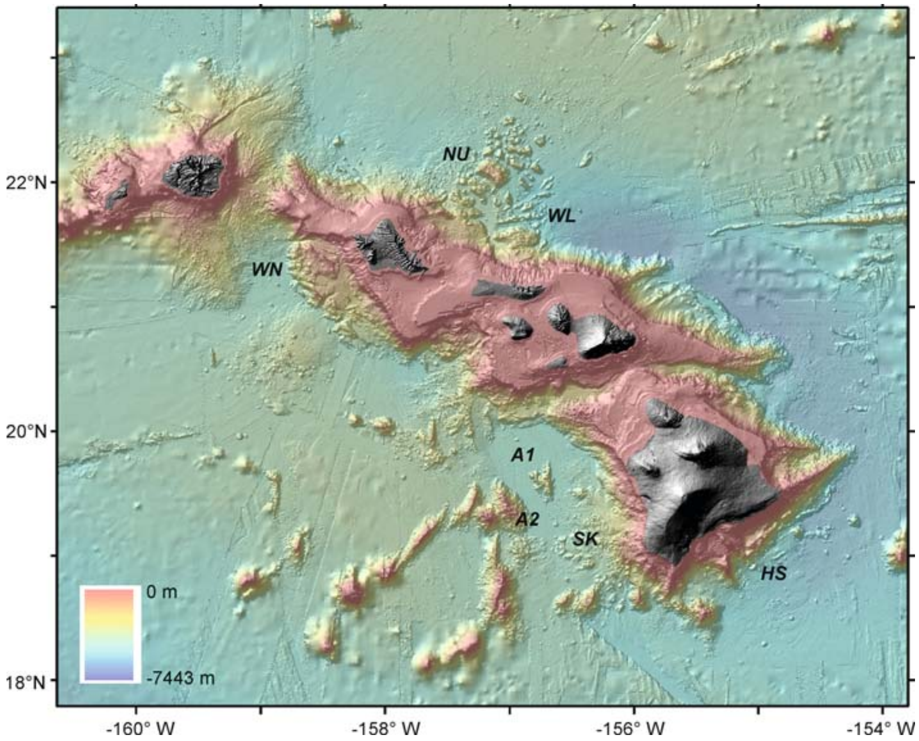
We see two principal reasons why monitoring submarine landslides and slow earthquakes associated with them is important. First, because catastrophic failure of submarine landslides can cause tsunami they represent significant hazards to coastal zones. Understanding and monitoring how slow slip may lead to accelerated slip and catastrophic failure is, therefore, very important in terms of hazard mitigation. Second, submarine landslide systems can be some of the most active as well as spatially confined deforming areas on earth and so they represent excellent targets of study for furthering our understanding of the general fault failure process. For instance a pertinent question for which we do not yet have an answer is: are fault frictional properties homogeneous enough that the occurrence of slow earthquakes on a detachment fault plane underlying a landslide could relieve stress on the fault or do the slow earthquakes in one region load a neighboring seismogenic patch bringing it closer to a large sudden failure (i. e. an earthquake)?

While installation and operation of GPS networks on land is now relatively routine and somewhat inexpensive, the in situ monitoring of submarine landslide motion represents a significant technical challenge with accordingly higher costs. Submarine geodesy e. g. [7], however, is a nascent and rapidly evolving field with relative and absolute positioning techniques being intensely studied and developed. The near future is sure to see many advances in our monitoring and understanding of submarine landslides and slow earthquakes due to the application of submarine geodetic techniques.

Introduction

Submarine Landslides

The last 30 years have seen a dramatic increase in the recognition of submarine landslides world-wide, due largely to the increased prevalence and capability of swath and side-looking sonar mapping systems and systematic submarine mapping programs. For instance, the side-scan sonar mapping of the Hawaiian exclusive economic zone in the 1980s resulted in the discovery that massive submarine landslides are spatially distributed along the entire Hawaiian Ridge [8] (Fig. 1). These features are some of the largest landslides on the planet, with more than 70 attaining lengths greater than 20 km and some having lengths greater than 200 km and total volumes exceeding 5000 km^3. Since then, submarine landslides associated with other volcanic islands e. g. [9,10,11], mid-ocean ridges [12], and continental margins [13] have also

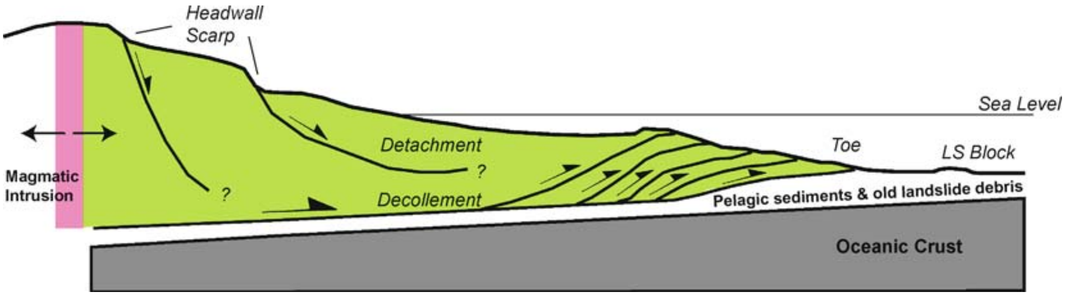


Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, Figure 1
Topographic and bathymetric map of the Hawaiian Islands (data from http://geopubs.wr.usgs.gov/i-map/i2809/). Studied submarine landslides are indicated: HS, Hilina Slump; SK, South Kona; A1, Alika 1; A2, Alika 2; WL, Wailau; NU, Nu'uanu; WN, Wai'anae

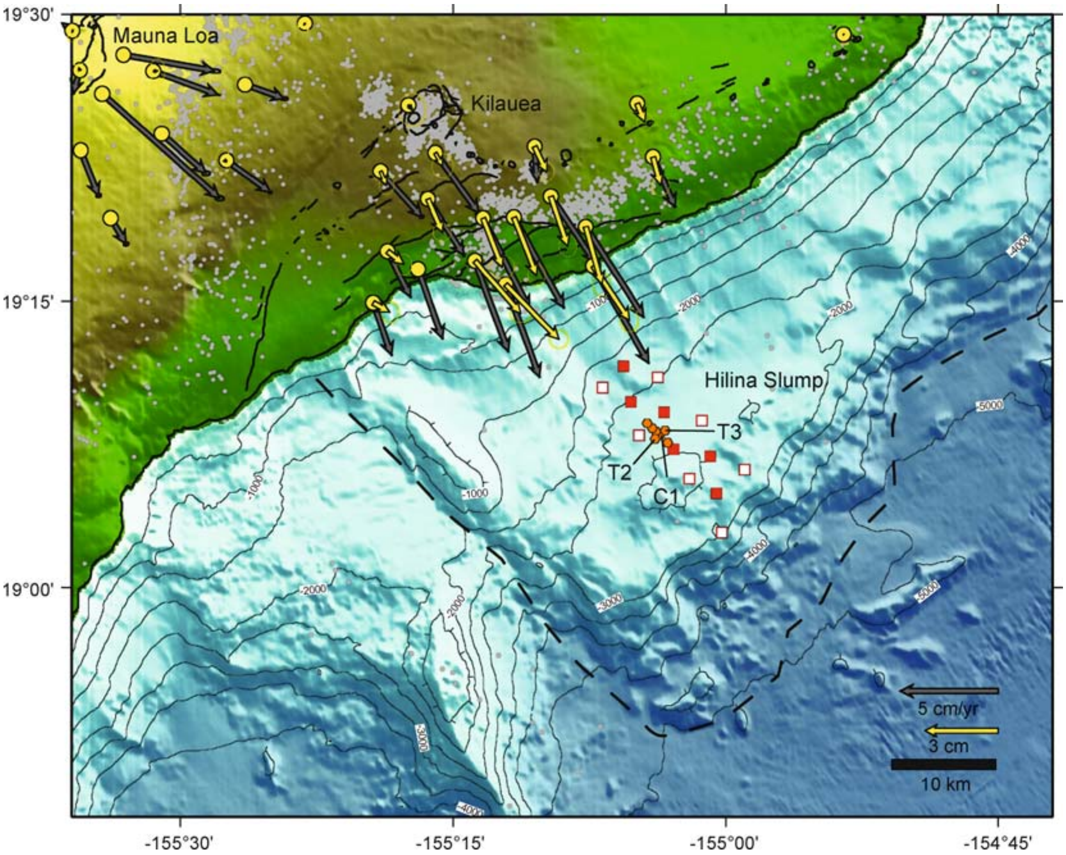
been studied. Of these, the slopes flanking volcanic islands, especially when they are in their steeper-sloped shield-building stage [14,15], tend to be particularly susceptible to landslide instability and so they have been the focus of much recent research in the Canary Islands [16,17] and especially in the Hawaiian Islands [18,19,20,21,22,23]. For instance, most of the Hawaiian submarine landslides are thought to be inactive, except for those on the flanks of the Big Island e. g. [21].

The morphology of submarine landslide features is similar to their subaerial counterparts. In map view they generally exhibit lobate and hummocky bathymetry. In cross-section, a wedge-shaped region of deformed material thins down-slope and is underlain by a gently-sloping planar dislocation surface (sometimes referred to as a basal detachment or 'decollement') separating the deformed carapace from the underlying undeformed substratum (Fig. 2). An upslope extensional head-scarp region transitions into a contractional fold belt towards the toe. The normal faults in the upslope region typically intersect the surface at high angles (> 45 degrees) and are manifested as scalloped-shaped scarps at the head of the slide that separate regions of differentially tilted fault blocks; in the sub-surface they may continue at high angles until they intersect the basal decollement, or they may sole with depth either into the decollement or into another sub-horizontal slip-surface [24]. The contractional regions are characterized by folded and bulging layers, closed depressions, and steep toes [20,21,25]. Moore et al. [8] separated the Hawaiian submarine landslides into two principal types: 'slumps' and 'debris avalanches'. The slumps are wide (up to ∼ 100 km), deep-seated (∼ 10 km thick), and have surface inclinations of up to 3 degrees while the debris avalanches are long (up to ∼ 230 km), shallowly-seated (50 m–2 km thick), and have surface slopes generally less than 3 degrees.

Concomitant with the mapping of the landslide features has been the increasing recognition that sudden submarine landslide movement can cause tsunami with destructive implications for coastal societies e. g. [26]. A particularly well-known example of this scenario is the 1929 Grand Banks failure [27]. Moreover, in the 1990s alone workers have attributed at least 5 tsunami events to catastrophic landslide failure sources: (1) 1992 Flores Island, Indonesia [28]; (2) 1994 Mindoro, Phillipines [29]; (3) 1998 Papua New Guinea e. g. [30]; (4) Kocaeli, Turkey [31]; (5) 1999 Pentecost Island, Vanuatu [32]. In Hawaii, the $M_w 7.7$ 1975 Kalapana earthquake, most likely



Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, Figure 2
Schematic cross-section of a submarine landslide flanking an active ocean-island volcano (after [21])

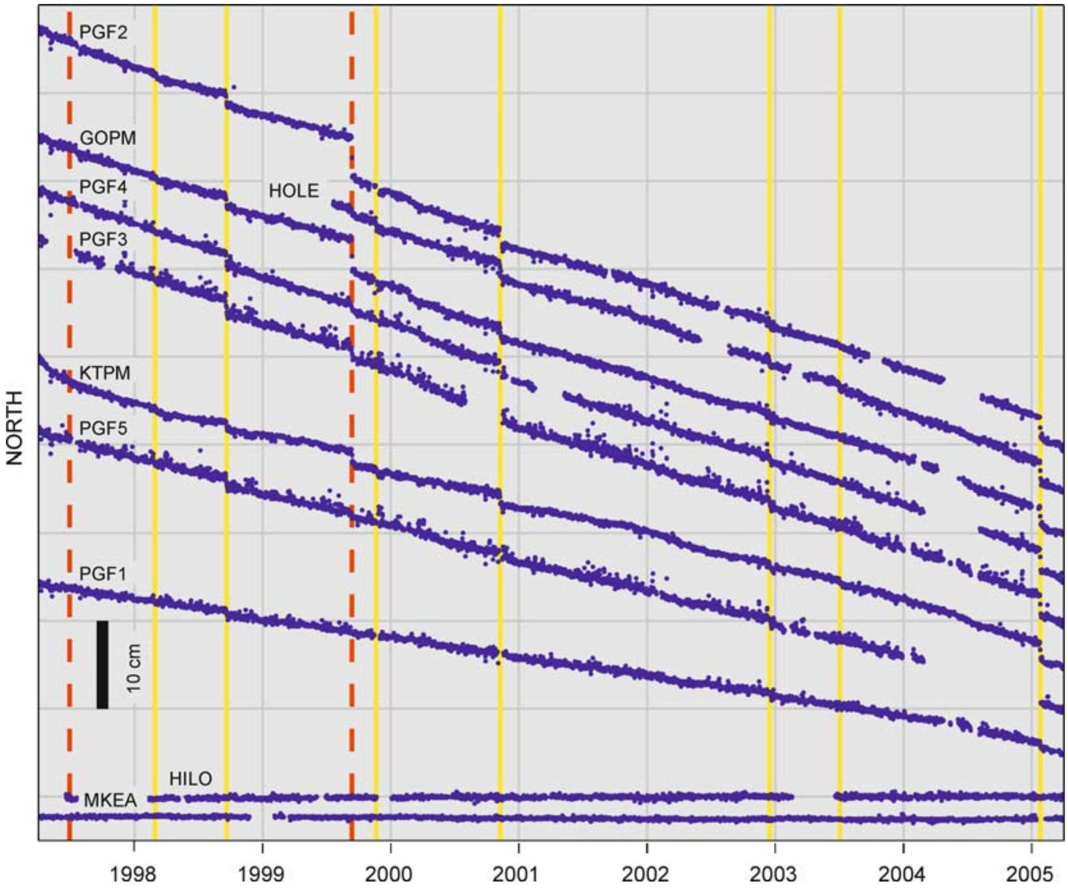


Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, Figure 3
Map of geodetic networks on the Hilina Slump overlain on topographic/bathymetric map. *Yellow circles* on land are CGPS sites operated jointly by the USGS Hawaiian Volcano Observatory, University of Hawaii, and Stanford University. *Red squares* offshore (*open and filled*) are seafloor geodetic sites operated by Scripps Institution of Oceanography. *Orange circles* offshore are acoustic extensometer sites deployed by our group with locations of transponders (T2,T3) and transceiver (C1) indicated. *Grey vectors* are average horizontal velocities from 1997–2005. *Yellow vectors* are horizontal motions from the January 2005 slow earthquake. *Grey dots* are earthquakes from the HVO catalog for the period May 2004–2005

due to slip of the fault surface underlying an active submarine landslide [33,34,35], caused local loss of life and damage in Southern California.

From a hazards standpoint it is particularly important to understand how tsunami are generated by submarine landslides because the propagation time between tsunami generation to runup is typically on the order of minutes. For instance, a catastrophic failure of the west side of the island of Hawaii would likely send tsunami waves around the Hawaiian islands that would reach the densely populated areas of Oahu's Waikiki beaches in less than an hour and more likely ∼ 30 min (G. Fryer, personal communication, 2007). Simulating waves generated by a sudden, chaotic disturbance of the seafloor, as expected from a submarine landslide, is quite complicated e. g. [36,37] and not all workers agree on approaches or results. Murty [38], however, stressed that parameters such as slide angle, water depth, density, speed, duration of the slide are second order, while instantaneously displaced volume is likely the most important parameter controlling tsunami generation.

Despite the increasing awareness of their hazard, little is known about how submarine landslides actually move, largely because of the challenge of installing instruments and retrieving data from the submarine environment. Moore et al. [8] recognized that while slumps more likely move relatively slowly, debris avalanches could be deposited very rapidly based on, for instance, uphill flow of material in the distal portions of the landslide deposits. In agreement with this, the estimated downhill velocities from the 1929 Grand Banks event was 60–100 km/h [39] whereas many studies have documented ∼ 6–10 cm/yr horizontal and vertical velocities associated with a submarine landslide flanking the Island of Hawaii's Kilauea volcano, the Hilina Slump [4,5,6,40,41,42]. Recently, GPS data from the Hilina Slump have elucidated that not only does the slump move at the above-stated, fairly smooth, background velocities, but also that the slump will occasionally deform in discrete accelerated cm-scale motions equivalent to $M6$ earthquakes but without the shaking [4,5,6]. These 'slow earthquakes' last hours and accommodate cm's of ground displacement (Fig. 3). It is not currently known, however, how the occurrence of a slow-earthquake in a submarine landslide system will affect its future movement by either making it more or less probable of failing catastrophically.



Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, Figure 4
Time series of north component of motion for selected GPS stations. Offsets of 7 slow earthquakes are identified by *yellow lines*. *Red dashed lines* are magmatic diking events

Slow Earthquakes

The term 'slow earthquake' has been used to describe a variety of transient aseismic deformation phenomena including slow precursor events preceding large earthquakes [43,44,45,46,47,48], afterslip following earthquakes [49,50,51], variable fault creep rates [52,53,54,55], certain subduction zone thrust events with unusually long durations and large amplitude tsunamis for their size [56], and discrete fault-slip events that do not produce detectable seismic shaking but are accompanied by ground displacements very similar to those produced during earthquakes [1,2,3,4,5,6,57]. Hereafter, when we use 'slow earthquake' we will refer to this latter description although the term was first used in the modern literature to describe a slow precursor to the great Chilean 1960 earthquake [43].

In contrast to slow precursor events whose ground motions, like traditional earthquakes, are measured in seconds or minutes, slow earthquake (SE) displacements usually accrue over time periods ranging from hours to days and so they have been typically sensed with geodetic methods. In the last decade, the proliferation of continuous GPS (CGPS) networks has led to numerous SE observations and the discovery of some very rich behavior. In certain regions SEs have occurred with very regular periods [2,4,58], they are often associated with non-volcanic tremor [59,60] and, apparently, SEs follow very different scaling laws (moment vs. duration) than traditional earthquakes [61]. Explanations for SE slip behavior has varied to date. For subduction zones, the combination of deep (> 35 km) SE sources and their association with tremor (a phenomenon initially thought to be caused by forced fluid flow [62] but more recently also explained in terms of shear failure [63,64]) led to one current hypothesis that SE mechanics are controlled by water released during metamorphic phase changes at the interface between a subducting and overriding plate [65,66]. Other explanations have invoked rate- and state-variable frictional behavior to suggest that SEs occur preferentially at transitions between velocity strengthening and weakening regimes on a fault plane [67,68,69] and that temporally varying climatic load changes could help explain SE periodicity [58].

Due, in part, to their large magnitude deformation signal and the high concentration of CGPS networks focused on them, subduction zones have dominantly been the location of the most SEs to date [1,2,57,59,70,71,72]. Recently, the CGPS network on Kilauea volcano's mobile south flank has recorded multiple SEs [4,5,6] (Fig. 4) and it is through the Kilauea events that SEs have come to be associated with submarine landslides.

Monitoring Motion: Subaerial and Submarine Geodetic Methods

Not surprisingly, much more is known about the motion of subaerial than submarine landslides. Geodetic measurements on land combined with contemporaneous measurements of other properties (pore-water pressures, strength of materials, etc.) have allowed, in some cases, a very thorough understanding of how landslide motion is related to driving forces such as gravitational stresses and rainfall. For instance, Baum and Reid [73] instrumented a slow-moving submarine landslide in Honolulu's Manoa valley with extensometers recording at 15 minute intervals and rain gauges and found a direct correlation between rain fall and deformation events. Similarly, Malet et al. [74] showed that GPS-measured surface velocities increased to as high as 20 cm/day following periods of higher rainfall during May 1999 at the Super-Sauze earthflow in the French Alps. At a slightly different scale, Hilley et al. [75] used InSAR to simultaneously map deformation of multiple landslides in California's Berkely Hills at $\sim$ monthly intervals and found that landslide motion correlated with times of high precipitation and that during the 1997–1998 El Nino event displacement rates doubled from the background rate of $\sim$ 27–38 mm/year, albeit with a $\sim$ 3 month time lag between the onset of motion and the high precipitation.

Much of the current knowledge that we have about the motion of submarine landslides comes from their easier-to-monitor subaerial portions, such as at Kilauea's Hilina slump (Fig. 3). In the case of the Hilina Slump, fully 3/4 of the feature resides offshore at depths greater than 2000 m and the deformation monitoring network is necessarily concentrated on the down-dropped blocks near the headwall scarp of the entire system. Recently, however, submarine methods have started to provide geodetic information from the ocean floor itself e. g. [7], e. g. [76,77,78,79]. As these methods become more cost-effective and widespread they will surely yield much insight into submarine landslide kinematics and, eventually, be employed in operational hazard monitoring/mitigation scenarios.

Whether a network is solely subaerial, submarine, or some combination of the two, it is important to consider that monitoring strategies can vary substantially depending on the time duration of the expected signal, the desired threshold of detection, and the desired time latency for individual solutions. For instance, simple detection and warning of catastrophic landslide failure needs the most rapid solution latency but coarse detection thresholds (meter rather than millimeter level, for instance) are appropriate. Conversely, if the goal is to detect small, potentially precursory motions such as slow earthquakes, then detection threshold must be as sensitive as possible along with solution latency being low.

Subaerial Geodesy: GPS

A multitude of geodetic techniques, from mechanically- to electromagnetically-based, are currently employed on landslides to measure motion caused by a range of deformational phenomena spanning opening of small surface cracks to the motion of kilometer sized blocks e. g. [74]. Data from Global Positioning System (GPS) networks has increasingly contributed to the library of observations associated with landslide motion. In particular, for submarine landslides that are large enough so that their motion causes earthquakes or slow earthquakes [4,6,80], a technique which is suitable for inter-station distances measured in kilometers is most appropriate and so below, we concentrate on the use of GPS with submarine landslides.

GPS GPS networks capable of sub-cm to mm-scale 3-dimensional ground motion detection are now deployed in many of Earth's most actively deforming zones (see for example, http://sps.unavco.org/crustal_motion/dxdt/) and readers are referred to thorough reviews of the general technique and its application for geodynamic studies [81,82,83].

Crustal motion GPS networks are usually divided into two types: those that record data continuously (CGPS) and those whose individual monuments are occupied less frequently in survey mode (SGPS). Depending on a variety of factors including the modernity of the receiver, the bandwidth of telemetry networks (should they exist), and the storage capacity of the archival center, CGPS sampling rate generally varies between once every 30 and 1 s (though most modern receivers are capable of sampling at frequencies higher than 1 Hz). SGPS sampling is more varied though it usually comprises re-occupation of sites at intervals ranging from months to years with occupation times of hours to days and sampling rates similar to CGPS. Accordingly, SGPS networks are more useful for wider ranging spatial characterization of deformation phenomena rather than for the rapid detection of motion or for tracking temporal evolution during transient events.

High rates of sampling alone, however, do not guarantee that CGPS network positional solutions will achieve their highest precision and/or accuracy. Assuming that all sites within a CGPS network have stable monuments and high-grade geodetic antennae and dual frequency receivers, the most important components of its error budget for deformation monitoring are: (1) integer ambiguity resolution; (2) orbital estimation; (3) atmospheric delay estimation; (4) antenna multipath; (5) satellite constellation geometry, and (6) intra-network baseline length. For networks monitoring landslides with spatial scales on the order of kms or tens of kms, however, the baselines are short enough that errors scaling with baseline lengths are small contributors to the error budget. In addition, as absolute positioning in a global reference frame is not essential, precise orbital estimation is less important. For the other error sources, freely available software packages such as GAMIT, GIPSY, and BERNESE (http://facility.unavco.org/software/processing/processing.html), and precise orbit processing centers such as the IGS (International GNSS Service, http://igscb.jpl.nasa.gov/) or the Scripps Orbit and Permanent Array Center (SOPAC, http://sopac.ucsd.edu/) usually allow good enough mitigation of these errors so that daily GPS solutions have resolution on the order of 2–5 mm in the horizontal and ~ 2–3 times worse in the vertical e. g. [83]. This resolution rule-of-thumb, however, generally applies to post-processed data with occupation times exceeding ~ 6–8 h, use of precise orbits, and the best atmospheric estimation techniques e. g. [84]. Because of the hours-to-days delay needed for estimating various grades of precise orbits and the computational time needed to estimate all the parameters for all sites with this much data this standard rule-of-thumb cannot necessarily be applied to a real-time or near real-time solution.

Real- and Near-Real Time GPS Processing There are a variety of processing techniques that may be suitable for real-time or near-real time monitoring applications for the subaerial counterparts to submarine landslides. For instance, the well-known real time kinematic (RTK) positioning technique frequently employed by the surveying community applies differential corrections sent via radio link between stations to yield site position estimates with cm-scale precision over baselines up to ~ 10 km in real-time [85]. The technique works best for baseline distances less than ~ 10 km typically, because the assumption of correlated errors between stations is not necessarily valid and differential corrections cannot be accurately applied for larger baselines. Thus, RTK may be a suitable monitoring technique if a more spatially contained portion of a landslide, such as a fault zone, is a particularly good indicator of motion for the overall larger unit.

It is the number of measurement epochs required to resolve the integer-cycle phase ambiguity inherent with GPS data, however, that is the principal factor limiting the temporal latency of high resolution GPS positioning solutions. Kinematic techniques typically require minutes worth of GPS data (when sampled at 1–30 s) in order to resolve integer ambiguities for initialization and reinitialization if a cycle slip or loss of phase lock occurs during measurement, although progress is being made on mitigating ionospheric effects and reducing time of integer ambiguity resolution e. g. [86]. Regardless, if GPS station locations are chosen with relatively good sky-view then loss of lock due to poor satellite visibility should be minimal.

Recently, Bock et al. [87] developed a method of resolving integer ambiguities from a single epoch of dual frequency GPS phase and pseudo range data that provides independent epoch-by-epoch position estimates over baselines as large as ∼ 40 km. Their method allowed instantaneous positioning resolution of 1.5 cm in the horizontal and 7–8 times worse in the vertical coordinates. Langbein and Bock [88] applied the technique to determine offsets in positional time series due to slip events recorded by the Parkfield, California CGPS network which has similar spatial scales to a typical large submarine landslide. They found that offset sensitivity was ∼ 5 mm for a 2 s sampling window and this decreased to ∼ 2 mm when a 60 s window was used [88]. To the best of our knowledge this represents the current state-of-the art in terms of real-time detection of motion over spatial and temporal scales typical of submarine landslides.

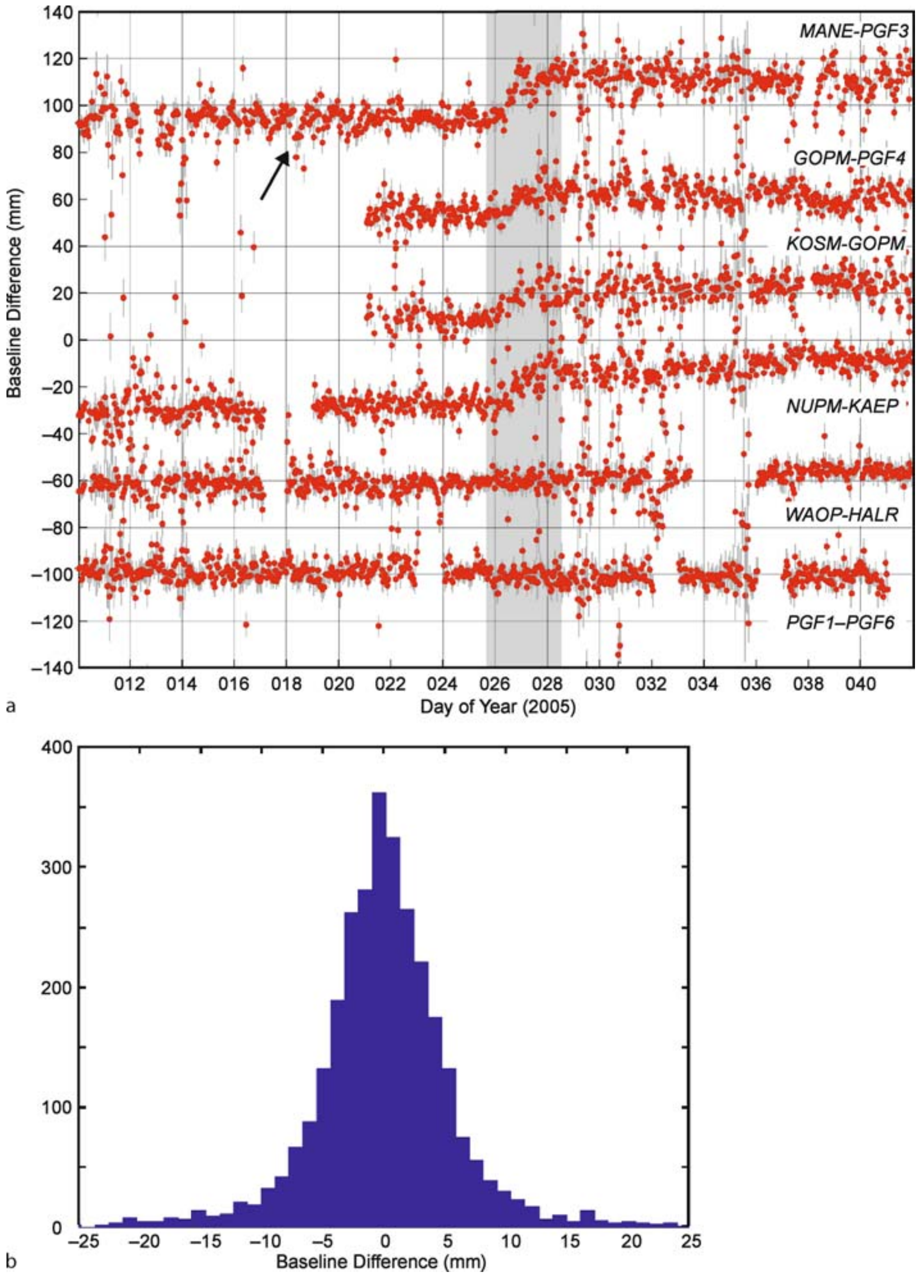
At Kilauea, we have developed an automated *near* real-time processing strategy for our monitoring efforts at the Hilina Slump (Figs. 3 and 5) [89]. The CGPS data are telemetered hourly by radio modem back to the USGS Hawaii Volcano Observatory (HVO) where they are retrieved via FTP. We use a sliding window approach, collecting all data available within the most current two-hour period and performing a network solution using the PAGES [90] processing software. We use the 30 s, ionosphere free phase combination as the observable and use and hold fixed the IGS ultra-rapid orbits and apply the NGS antenna phase calibration patterns for each site [91]. We also apply the IERS standard solid Earth tide [92] and Schwiderski ocean tide loading corrections [93,94]. Every half hour we estimate a piece-wise, linear neutral atmosphere (troposphere) correction and one set of N–S and E–W neutral atmospheric gradient corrections for each site. We apply a 'weak' constraint of 10 cm to the atmospheric corrections based on examination of previous adjustments.

Figure 5a shows an example of baseline change time series from indicative stations derived from this approach and re-run in a simulated near real-time mode for the time period bracketing the slow earthquake from January 26–28, 2005 at Kilauea. Generally, the hourly baseline change noise levels are on the order of ±10 mm, although because of some particularly large outliers, the standard deviation of the baseline changes (not including the time period of the SE) is closer to 20 mm (Fig. 5b). Some of the large excursions in the time series not associated with the SE are due to the high amplitude, strongly spatially and temporally varying atmospheric water vapor gradients often associated with tropical islands such as Hawaii e. g. [95]. For instance for the excursion in the MANE-PGF3 baseline in the middle of day 18 (Fig. 5a), the time series takes a sharp upward bend until the start of day 19 when it again begins to oscillate about its mean value from present days. This baseline change gradient is very similar to the onset of the SE near the beginning of day 26. It is clear from the remainder of the time series after day 27 that the ∼ 25 mm offset is permanent, however, and so indicative of a real deformation event.

In a real-time monitoring scenario, because of the 2 h temporal latency and the atmospheric noise levels on the order of 1/3 the maximum signal levels in the baseline change plots, it would probably require on the order of 12–24 h before this event could be definitively classified as a deformational event. This could certainly be useful to hazard mitigators for being aware that a slow event such as an SE was occurring, but not in the event of a rapidly accelerating catastrophic collapse where the detection and warning time must be on the order of minutes. For more rapid warning, more sophisticated filtering techniques could be employed (such as monument motion, and network-wide coherence assessment as described below for the Network Inverse Filter [96]) although regions of large atmospheric gradient will always be hampered by low signal-to-noise ratios unless more sophisticated atmospheric mitigation techniques such as tomographic mapping [97] are employed in a real-time manner. Accordingly, the small detection threshold levels of the epoch-by-epoch processing at Parkfield described above [88] must be taken in the context of the relatively low atmospheric delay environment present there.

Submarine Geodesy

Subaerial geodetic methods have, until now, provided essentially the entirety of geodetic evidence for submarine landslide-related deformation including slow earthquakes. Our understanding of submarine landslide motion and slow earthquake mechanics stands to increase dramatically in the coming decade, however, owing to the rapidly advancing field of submarine geodesy which can now reliably provide in situ measurements of seafloor deformation. The submarine environment is particularly challenging for geodesy techniques especially if they require the propagation of energy across a medium (ocean water) that



Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, Figure 5
a Baseline difference (position – median position) and 2σ errors for near-real time processing of selected sites at the Hilina Slump during early 2005. See text for description of processing. *Grey shaded region* indicates the time interval of the slow earthquake. *Black arrow* indicates an example of an anomalous trend due to atmospheric delays discussed in text. b *Histogram* of baseline difference values for all of the sites in a, excluding the time period of the slow earthquake

can exhibit highly spatially and temporally variable material properties.

Direct Path and Indirect Acoustic Approach One of the submarine geodesy techniques most frequently employed to date is direct path acoustic measurement of baseline length changes using acoustic transponders. This method has been used to make in situ observations of tectonic motions on the seafloor at the Juan de Fuca ridge where one study found cm-scale motion occurring over a period of a few days associated with an on-axis eruption [77] and another found no detectable motion over a period of a few years at a quiescent ridge segment [76]. By interrogating each other, pairs of instruments use two-way travel time measurements to constrain the inter-station distance. After correcting for variations in sound speed that result from changes in water temperature, salinity, pressure, and local tilt of the monument itself, these systems yield sub-cm precision for individual measurements. Operating frequencies typically range from 7.5 to 108 kHz and in deeper isothermal waters, the upward refraction of sound waves prevents signals from one transponder being received by the other at distances greater than ∼ 1 km [79]. Travel-time is determined simply by correlating the transmitted and received signals and picking the peak of the resultant correlogram. In quiet operating environments this can usually be done at the scale of ∼ 5 µs, resulting in a ∼ 4 mm range error [7].

For the greater depth ranges, Sweeney et al. [79] devised an approach that allowed ±2 cm resolution measurements for baselines up to 10 km at 2500–2600 m depths at a stable site on the Juan de Fuca plate. Sweeney et al. [79] suspended an acoustic interrogator hundreds of meters above the seafloor in a position acoustically visible to an array of seafloor-mounted instruments. In a manner similar to Spiess [98] they estimated the relative positions of the stations by moving the interrogator and collecting acoustic range data multiple locations.

The focus of these initial deployments of the direct and indirect acoustic approach systems was on measuring annual tectonic rates, and not necessarily on capturing transient events such as SEs that occur over hours or days. Accordingly sampling rates in current seafloor geodesy projects typically do not exceed a few times per day, largely because providing adequate power to seafloor instruments is still financially prohibitive. However with cabled oceanographic observatories scheduled to come online in 2007–2010 (http://www.neptunecanada.ca/; http://www.orionprogram.org/) the power delivery problem could be solved at high priority seafloor sites.

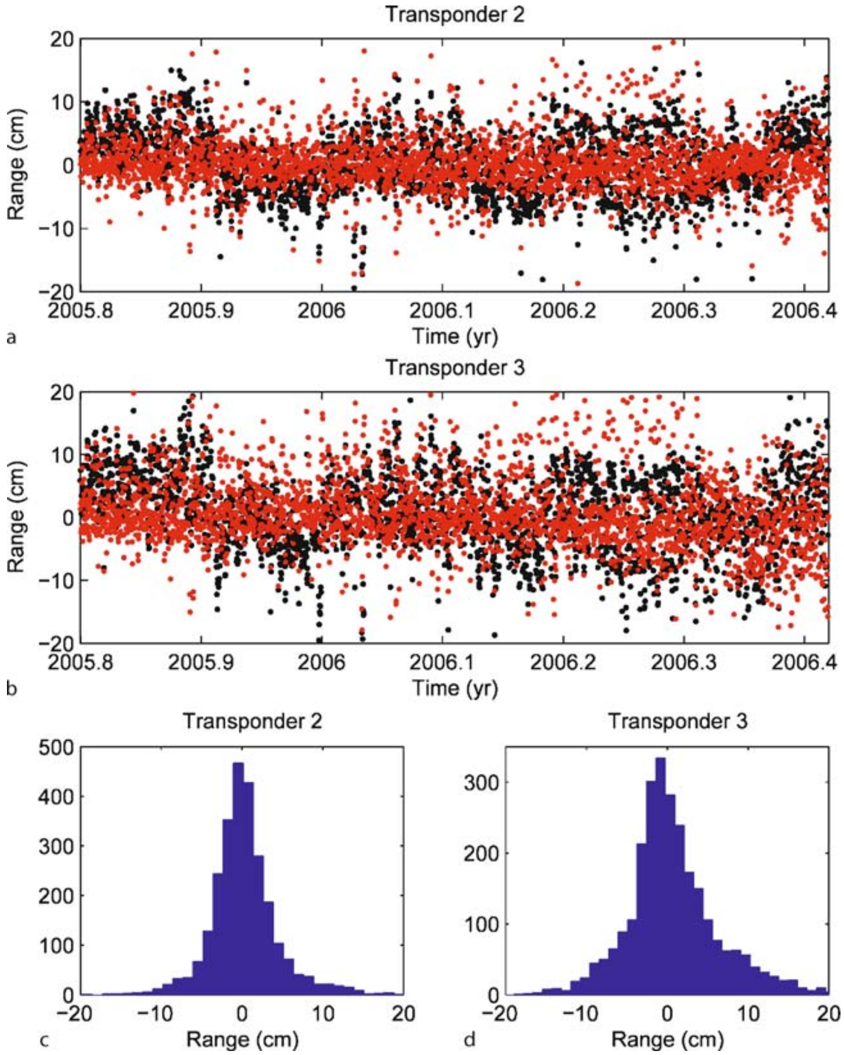
From October, 2005 through June, 2006 we deployed seven 10 kHz Linkquest transponders mounted ∼ 3 m above steel tripods and spaced over a distance of ∼ 3 km on the Hilina Slump (Fig. 3). One of the goals of this pilot project was to evaluate the performance of the transponders for use in submarine landslide monitoring at significant depth; each unit operated at a depth between 2640 and 2690 m. Battery power restrictions for the ∼ 8 month duration of the project meant that the instruments ranged to one another 12 times per day.

In Fig. 6 we show range change time series from two transponder-pair baselines (∼ 530 and 683 m respectively) for which data recovery was complete over the experiment's duration. For the other transponder pairs data recovery was not as complete because either: (1) the unit was knocked over by local mass-wasting events or (2) the baseline distance was too long and data dropouts occurred when the transponders lost sync with one another. The two way travel times were picked from the peak of the correlation function between the outgoing and received waveforms at one end of the baseline. The black dots in Fig. 6a and b show the raw measurements, while the red dots show the distance measurements corrected for the variations in sound speed due to temperature changes, which were measured by an external conductivity and temperature sensor mounted on the transponder frames. The raw time series show significant long-term trends due to temperature variations that are effectively removed using just the temperature measurements at the transponders. We did not correct for salinity variations because all of our conductivity sensors provided contaminated data due to clogging by the local mass wasting events.

For these baseline pairs which were oriented approximately perpendicular to the maximum expected motion of the Hilina Slump we expect essentially no motion for such short baselines and during such a short measurement period. At the 1σ level individual measurement noise is ∼ 4.1 and 5.7 cm respectively, though it is clear that smaller, cm-scale changes would be detectable given a long enough time period. For instance, daily estimates of baseline length have approximately a 1 cm standard deviation. For rapid event detection, baseline changes would need to be on the order of 10 or more cm or measurement frequency would need to be increased.

GPS-Acoustic Method The acoustic measurements described above provide only relative measurements of baseline length changes; however, for submarine landslide monitoring efforts it may be desirable to place submarine and subaerial measurements in a single regional or global reference frame. Largely with the aim of furthering seafloor active tectonics studies, a research group spearheaded by the late Fred Spiess conceived of, and have begun implementing integrated GPS and acoustic seafloor geodesy (GPS-A) studies, described in great detail by Spiess et al. [7] and references therein.

Briefly, GPS-A positions seafloor geodetic markers in a global reference frame by combining kinematic GPS positioning of a sea-surface platform (ship or buoy), precision underwater sound travel time measurement over km-scale path lengths, and a strategy for eliminating the large



Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, Figure 6
a,b Baseline difference (position – median position) for transponder-transceiver pairs (T2-C1, T3-C1 in Fig. 3) pairs from October 2005 – June 2006. See text for processing description. ***Black dots*****, raw baseline difference uncorrected for temperature.** ***Red dots*****, baseline differences corrected for temperature. c,d** ***Histogram*** **of baseline difference values from a and b, respectively**

errors in travel time measurement arising from a spatially and temporally varying sound speed structure in the near surface portion of the water column. The error mitigation strategy is based on the recognition that the location of the sea surface midpoint above the center of a triangle of seafloor transponders (themselves positioned via a near-bottom acoustic survey) is the point at which the three sea surface-to-seafloor travel times are identical and so, independent of sound speed.

Spiess et al. [7] reported repeatabilities of ±0.8 cm and ±3.9 cm in the north and east components, respectively, of seafloor measurements on the Juan de Fuca plate from 1994–1995. More recently, Gagnon et al. [78] reported velocities with ~ 5.5 cm magnitudes and 0.6–0.8 cm 1σ errors from surveys between 2001–2003 on the updip portion of the Nazca-South America subduction zone in the Peru trench.

Optical Path Length Zumberge [99] combined commercial subaerial surveying technology (an EDM) with an optical fiber strainmeter in order to devise a low-power, cost-effective system for the harsh seafloor operating environment. EDMs (electronic distance meters) typically can measure distances with 1–2 mm precision over several km by transmitting intensity modulated infra-red energy and measuring travel time from far away reflective surfaces. Zumberge [99] modified the EDM for seafloor geodesy by focusing the transmitted beam into the fiber's core and gathering the reflected light from the fiber's far-end into the EDM receiver optics.

The composite instrument comprises an optical fiber (enclosed in a hermetically sealed stainless steel casing) that stretches between two anchors, one 'active' and one 'passive', separated by several hundred meters on the seafloor. The active anchor houses the EDM and all supporting electronics and the passive anchor's purpose is to fix the fiber to the seafloor. In a test of their instrument, Zumberge et al. [100], report 1mm scatter of distance measurements over a 500-m-long fiber during a 50 day long period.

One of the obvious advantages of the optical path length technique is the high precision, isolation from sea water-borne errors, and the continuous nature of the measurement. The disadvantage, however, is that the technique is limited to short baselines ($< \sim 1$ km) and so the precise location of straining regions must be known a priori. Nonetheless, for monitoring purposes with very well-defined targets such as the headwall portion of an active landslide, the technique holds much promise.

Pressure Sensor Vertical Deformation Phillips et al. [40] recently developed a new technique that used pressure sensors in campaign-mode repeat surveys to measure vertical deformation rates of the seafloor at depths exceeding 2000 m at the offshore portion of the Hilina Slump. Depth and pressure are related to one another in a straightforward manner through the hydrostatic equation, though for measuring seafloor deformation rates at the cm/yr level many other reductions must be performed (Phillips et al. [40] – also her thesis). The pressure sensor method is technologically challenging, requires much ship time, and requires during each site visit both a geodetic monument and a pressure sensor (lowered from a ship) to be operating in close proximity to one another at depth on the seafloor. Because, in each revisit it was not practical to collocate the pressure sensor in a repeatable fashion on the benchmark, Phillips et al. [40] found it necessary to measure the vertical offset between the pressure sensor and the benchmark via acoustic ranging.

Important error sources for this technique are poor knowledge of α the specific volume of the water column (the reciprocal of density) and the speed of sound in depth. Due to changing tides, secular and seasonal barometric changes it is also necessary to utilize a common, stable reference, such as mean sea level (MSL) which, in turn, requires estimation of the geopotential anomaly.

Ultimately, Phillips et al. [40] conclude that their data show significant vertical deformation (9 ± 2.4 cm/yr) in the mid-section of the HS and negligible deformation towards the outer bench (just inboard of the toe of the entire landslide feature). These data are consistent with dislocation models delimiting the potential amount (25.0–28.1 ± 7.3 cm/yr), spatial extent (24.8–27.0 ± 0.5 km seaward of Kilauea's East rift zone), and depth (7 km) of the principal slipping surface below the landslide. This is a significant advance not only because it represents the first data set that allows a glimpse of how strain is partitioned in a massive submarine landslide such as the Hilina Slump, but also because it represents an ongoing monitoring project at a submarine landslide, albeit at temporally sparse sampling rate.

Data Analysis and Inversion

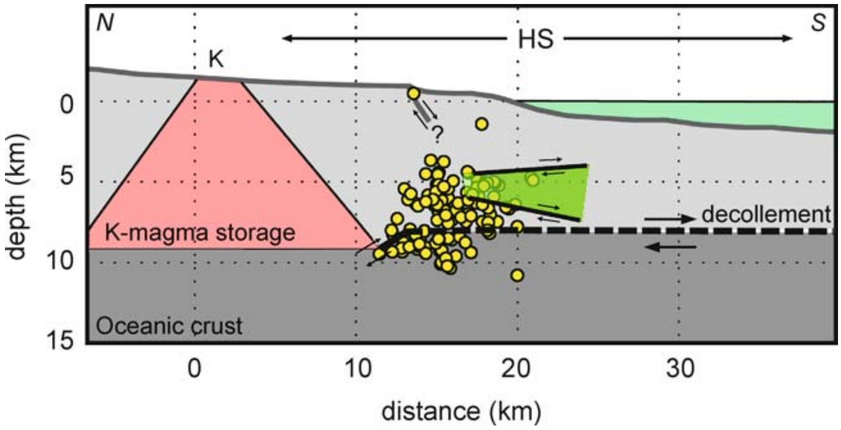
In addition to the challenge of acquiring geodetic data from submarine or even subaerial landslides is the added challenge of inferring sub-surface fault geometry and/or deformation processes, with attendant realistic error bounds on estimated parameters, from a data set of earth surface displacements. We identify two principal types of analysis modes: (1) process-based and (2) hazards-based. Process-based analysis focuses on deriving the most accurate and, if possible, complete assessment of the factors involved in the observed landslide motion. Hazards-based analysis has two primary components: (a) rapid warning and (b) long-term hazards estimation. Rapid warning comprises, with the smallest temporal latency possible, providing information regarding the current state of the landslide system and how it relates to current, short- and long-term dangers to life and infrastructure. Long-term estimation comprises making probabilistic statements about the components and potential future behavior of the system and so temporal latency is not a limiting consideration.

For either of the two types of analysis modes, the first objective is to usually try to relate the measured displacements d, of monuments at the surface of the earth to the source parameters, m, of the buried feature causing the deformation (usually considered a fault) through $G(m)$, a model of the deformation process. For instance, one very common model formalism, the dislocation in an elastic half-space, is a non-linear problem whose model vector m contains 9 parameters describing the location, orientation, and relative displacement of the dislocation approximating a faulting source [101,102].

Determining the most likely components of m and associated errors falls under the wide category of geophysical inversion for which there are a multitude of techniques e. g. [103]. In addition to the computational power available and the total amount of time allotted to the inversion, the choice of inversion method depends upon the nature of the model (for instance, linear vs. non-linear), the number of parameters in m, and the computational cost associated with individual model realizations and misfit assessments. Furthermore, it must be decided if the goal of an inversion is to obtain a rapid solution (a solution that satisfies some predetermined misfit level), an optimal solution (i. e. one 'best-fitting' solution), or rather to probe parameter space as thoroughly as possible with the goal of estimating posterior probability densities for each of the parameters for their value as indicators of resolution and uncertainty e. g. [104]. For instance, the most robust inversion method possible is the direct or 'grid' search, where the misfit for every possible permutation of m in parameter space is calculated. For more complicated or non-linear $G(m)$ however the computational cost associated with each forward model run can make a grid-search time prohibitive, even for a process-based analysis. For these types of more difficult problems, inversion based on Monte Carlo sampling. which collects pseudo-random samples from multidimensional parameter space as a proxy for the problem's true posterior probability density, $\sigma(m)$, has been found to be quite successful [105,106]. Indeed, it is common practice in the tectonic geodesy community to use the Okada model combined with some type of Monte-Carlo method to derive parameters of earthquakes and slow earthquakes from surface displacements observed with GPS [80,104,107].

The recent evolution of the work on the SEs at Kilauea demonstrates how important the inversion results are to the overall analysis either from a process-based or long-term hazards-based perspective. In their initial recognition of the November 2000 SE, Cervelli et al. [5] used a simulated annealing optimization routine [80] and the Okada model to invert the GPS observations and conclude that the most likely dislocation source for the SE occurred on a gently landward-dipping thrust fault plane at 5–6 km depth. Combined with the fact that the SE post-dated by 9 days a burst of local rainfall of nearly 1 m and reasonable estimates of local hydrologic parameters, Cervelli et al. [5] suggested that the increased pore pressure due to deeply percolating rainwater triggered SE motion by inducing a ~ 2 MPa pressure decrease of the effective normal stress on their preferred fault plane. Brooks et al. [4], however, showed from Gibbs Sampling inversion [104] results of GPS data on three additional SEs that posterior distributions of estimated fault parameters allow a wide family of equivalently plausible solutions, ranging from deeper seated decollement solutions at ~ 8 km depth to the more shallow fault plane favored by Cervelli et al. [5] (Fig. 7). Moreover, they showed that the other SEs were not associated with anomalous rainfall. Segall et al. [6] then used the same inversion method of Cervelli et al. [5] for the GPS data of the additional SEs relocations of high-frequency earthquakes triggered by the January 2005 SE, seismicity rate theory [108], and Coulomb stress modeling [109] to conclude that the SE (and other similar events) likely occurred at a depth of $\sim 8 \pm 1$ km on the main decollement plane below the Hilina Slump.

In this case, the crucial addition to the source inversion was the added constraint of the triggered microearthquakes which were relocated to depths near the decollement [6]. In their earthquake relocations, however, Segall et al. [6] did not use the full waveform data; rather, they used a double-difference-derived mapping with manual picks, assuming a 1D velocity model, between triggered events and previous high-precision relocations and tomography from elsewhere at Kilauea [110]. Two other studies, [111,112] performed high precision relocations using waveform cross correlation data and found different depths for the same cluster of events as Segall et al. [6]. Got and Okubo [111] suggested that their relocated events (including those triggered by the 1998 SE) do not illuminate a sub-horizontal fault plane but rather a deeper, steeply south-dipping reverse fault. Wolfe et al. [112] found that the triggered seismicity from the four SEs identified by Brooks et al. [4] consistently relocates on distinct map-view clusters aligned in the direction of the SE displacements themselves and in a subhorizontal band with depths of ~ 5 km. This is a solution consistent with Morgan et al. [21] who used seismic reflection data to identify moderately landward-dipping fault planes at similar depths. While the epicenters are well constrained in the Wolfe et al. [112] study, however, they were concerned that poor station geometry as well as near source velocity heterogeneity may bias the absolute depth of the relocations and the analyses may not be capable of distinguishing between shallow and deep fault zones. Thus, while the decollement is certainly the prime candidate on which a SE would occur (in agreement with the Segall et al. [6] suggestion), the analyses remain somewhat equivocal. This, in turn affects our best understanding of the most important hazard issue, the propensity of the decollement for a catastrophic failure. In order to more definitively answer these questions the above-mentioned research groups in conjunction with the Hawaiian Volcano Observatory of the US Geological Survey (the group charged with hazards analysis and



Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, Figure 7
North-south cross-section of the Hilina Slump after Got and Okubo [111]. K, Kilauea. HS, Hilina slump. *Black line* below K is the seismic (*solid*) and aseismic (*dashed*) position of the decollement from Got and Okubo [111]. Two *thicker black lines* with *green shaded area*, are the family of geologically plausible dislocation solutions equivalently supported by inversion of GPS data. *Yellow colored circles*, locations of microearthquakes from HVO catalog for ±5 days around each slow earthquake event

mitigation for the region) have teamed up in a joint seismological and geodetic research project aimed at definitively constraining the depths of triggered seismicity, and, hopefully, the depth of SE sources in the region.

While the above example focuses mostly on process and long-term hazards assessment, recent developments focused on deriving the time-dependent history of slip during a deformation event could also potentially be used for automated real-time inversion and event detection [96,113]. The Network Inverse Filter (NIF) [96] is a recursive Kalman filter algorithm that operates on either processed position estimates or raw phase data (rather than derived displacement rates) from an entire network of geodetic stations, includes a stochastic description of local benchmark motion, and finds a non-parametric description of slip rate on a fault plane as a function of time. In current implementations, the NIF employs Green's functions relating slip to surface displacement computed from the analytical solutions for a dislocation in an elastic half-space [101] although there is no reason other deformation models could not be employed [113]. Application of the NIF to the 1999 Cascadia SE allowed McGuire and Segall [113], for instance, to determine that slip rate on the slipping fault plane took up to ~ 20 days to reach its peak and that the southern portion of the fault had finished slipping before the northern portion began to slip. Cervelli et al. [5] and Segall et al. [6] also applied the NIF to the Kilauea SEs to derive source-time functions for the events. One caveat, however, is that the fault plane is not solved for by the NIF, rather, it must be known and held fixed a priori. For subduction zones with greater spatial geodetic coverage above the slipping portion of the fault this condition may be satisfied more satisfactorily than at submarine landslides.

Discussion: Slow Earthquake and Submarine Landslide Process

Largely because of CGPS subaerial measurements at Kilauea there are now a suite of observations of slow earthquakes related to submarine landslide motion. As it seems that SEs are a fairly general fault slip phenomenon, occurring at a variety of subduction zone locales globally, it stands to reason that SEs may be common to submarine landslide-related deformation, at least on the flanks of ocean island volcanoes. Clearly, as seafloor geodesy projects become more common we will learn if this is the case or not. It is also clear that the search for a theoretical understanding of the slow earthquake phenomena will be coupled to further understanding the motion of submarine landslides.

Currently, explanations for slow earthquake slip behavior focus on rock mechanic theory that follows an empirically derived rate- and state-variable frictional constitutive law based on laboratory experiments e. g. [114]. (The description below follows the summary and notation in Scholz [114]). The Dieterich–Ruina or 'slowness' law is expressed as:

$$\tau = [\mu_0 + a\ln(V/V_0) + b\ln(V_0\theta/\zeta)]\sigma \; ,$$

where τ is shear stress, μ_0 is the steady-state friction, σ is effective normal stress, V is slip velocity, V_0 is a reference velocity, a and b are frictional material properties, ζ is the critical slip distance and θ is a state variable that evolves according to:

$$\mathrm{d}\theta/\mathrm{d}t = 1 - \theta V/\zeta \,.$$

The frictional stability of the system, then, depends on $(a - b)$, the velocity dependence of steady-state friction, defined:

$$a - b = \delta\mu/\delta(\ln(V)) \,.$$

In the context of a simple spring-slider model approximating a slipping fault, the boundary between the stable and unstable frictional regimes will occur at a critical value of effective normal stress, σ_c, given by:

$$\sigma_\mathrm{c} = k\zeta / -(a - b) \,,$$

where k is the stiffness. When $(a - b) > 0$ the material is said to follow ‘velocity strengthening’ behavior and the system is stable – earthquakes cannot nucleate in this regime and earthquakes propagating into such regions will be abruptly terminated. When $(a - b) < 0$ the material is said to follow ‘velocity weakening behavior and the system is unstable for $\sigma \geqslant \sigma_\mathrm{c}$ – earthquakes will nucleate in this regime. When $\sigma \leqslant \sigma_\mathrm{c}$ the system exhibits oscillatory behavior and is said to be ‘conditionally stable’, it is stable under quasi-static loading but requires a discrete velocity perturbation in order for earthquakes to nucleate. Others interpret these relations slightly differently suggesting that nucleation occurs when matrix stiffness (scale and stress-rate dependent) drops to a critical value over a spatial scale large enough to promote rupture [68,115].

It may be in the boundary of this stability transition where slow earthquake slip behavior arises. For instance, a common observation of most of the subduction zone SEs is that they occur down-dip of the ‘locked’ zone or near the base of the seismogenic zone where earthquakes nucleate [1,57,58,69,71]. At these depths in subduction zones temperatures are close to the 450°C temperature at which feldspar starts to exhibit plastic behavior and so conditionally stable behavior would be expected [114], although it has recently been shown that transient oscillatory behavior may also arise naturally from system dynamics alone [116]. Faults in the conditionally stable regime, under steady-state loading, slip aseismically unless a perturbation to the system is large enough to push the fault across the stability boundary triggering an earthquake [117]. If the perturbation, however, is not quite large enough to push the fault across the stability regime then a period of stable sliding at increased velocity will occur (i. e. a slow earthquake) as the fault evolves back to steady-state [117]. It is not yet clear, however, how the explanations for subduction-zone related SEs may translate to the submarine landslide environment. While many subduction zone SEs apparently occur down-dip of the ‘locked’ zone where earthquakes nucleate along the subduction interface, at Kilauea, SEs apparently occur up-dip of the zone where the majority of earthquakes occur [4,6].

Recent studies have more explicitly focused on helping to explain SEs in terms of the rate- and state-variable formalism. For instance, Kato [67] used rate- and state friction to simulate the effect that ‘asperities’, velocity weakening regions surrounded by velocity strengthening regions, have on slip behavior and found that episodic SEs occur when the velocity-weakening patch size is close to the critical size of earthquake nucleation. Similarly, Liu and Rice [68] found that when they applied along-strike variations in the frictional $(a - b)$ parameter in models of subduction zone processes, that SE-type transient deformation events emerged spontaneously near the downdip end of the seismogenic zone. They suggested that the downdip end of the seismogenic zone is likely to be in the conditionally stable boundary between unstable- and stable regimes and that this could allow the SE behavior. More recently, Lowry [58] employed further theoretical implications of rate- and state-variable friction under resonant loading conditions [115,118] to suggest that Earth’s response to climatic redistribution of atmospheric, hydrospheric, and cryospheric loads could lead to resonant fault slip behavior that could explain observed periodic slow earthquakes on the Cocos-North America plate boundary at Guerrero, Mexico.

Shibazaki and Shimamoto [119] recently proposed an alternate model of slow earthquakes that was motivated by laboratory experiments on the velocity dependence of frictional stability. They introduced a cutoff velocity to the rate-state formulation to mimic laboratory results where fault surfaces transition from velocity weakening at low sliding velocities to velocity strengthening at higher velocities. Implementing the laboratory values for these transitions reproduced the slip (10^{-7} m/s) and rupture propagation velocities (km/day) seen in subduction zone environments. Their model predicts a linear relationship between these two velocities that can in principle be inferred from high quality continuous geodetic measurements of slow earthquakes. This relationship may be an important way to relate observations of slow earthquakes to mechanical models, particularly in regions such as Kilauea where the fault planes are shallow enough (< 10 km) to allow the details of the rupture to be resolved with high quality instrumentation. We note additionally that other recent model parameterizations [116] yield recurrence intervals and propagation velocities similar to observed events and Lowry [58] also suggested means of relating geodetic observations and model parameters.

Two additional factors that may modulate the frictional properties of faults in submarine landslide environments include local hydrologic and magmatic forces (in regions of active volcanism) [23]. For instance, the Nov. 2000 SE at Kilauea post-dated by 9 days an intense rain storm of nearly 1 m at the southeastern Big Island [5]. Cervelli et al. [5] estimated permeability using regionally appropriate hydrologic parameters for porosity and fault zone diffusivity and suggested that the 1 m of rain could have triggered SE motion by inducing a ~ 2 MPa pressure decrease of the effective normal stress on a gently landward dipping fault at ~ 5 km depth. Brooks et al. [4], however, showed that 3 other SEs at Kilauea were not preceded by anomalous rainfall and that other periods of anomalous rainfall were not accompanied by SEs. Additionally, prior to these events Iverson [23] was skeptical of rainfall triggers as he found from a rigid wedge analysis that, in order for ground water head gradients to be large enough to destabilize the Hilina Slump, implausibly large clay layers ($\sim$ 200 m thick) or very low hydraulic diffusivity ($\sim 10^{-11}$ m^2/s) needed to be present. Others, however, have found that magmatic injection-induced mechanical and thermal pressurization of fluids may help to explain Canary and Cape Verdes Islands flank instabilities [120]. Excess shear stresses exerted on a basal decollement because of rift zone magma injection have also been shown to be of sufficient magnitude to potentially cause slip [22,23]. Although others have suggested from geodetic observations of discrete dike events at Kilauea's rift zones that the dikes are injected passively, as a response to decollement slip, rather than as a trigger for it [121,122].

Future Directions: Slow Earthquakes and Submarine Landslide Monitoring

It is clear that expanded seafloor geodetic monitoring of submarine landslides would be extremely important in bettering our understanding of both the slow earthquake process and the hazards associated with submarine landslides. For instance, at the Hilina Slump, it is not even known how much of submarine portion of the landslide actually displaces the seafloor during a slow earthquake. Given the logistical and financial challenges associated with seafloor geodesy, however, it is reasonable to ask if it is worth it to society to monitor submarine landslides. In the Hawaiian Islands, for instance, one compilation suggests that there have been at least 6 tsunami-generating landslide events in the past 300 000 years [123]. For comparison, the average 50 000 year recurrence interval for such events is 1 to 2 orders of magnitude larger than typical earthquake recurrence intervals in Southern California where substantial resources are focused on earthquake-cycle related monitoring (http://www.wgcep.org/). As discussed above, however, landslide-generated tsunami can be quite damaging both locally and regionally. One study simulated $\sim$ 30 m wave heights reaching the California coast within 6 h of a massive collapse of the Hilina Slump at Kilauea [124]. Clearly minimizing the impact of such an event would be of societal benefit.

Continuous submarine landslide deformation monitoring, although expensive, is not out of the realm of possibility, as costs for seafloor geodetic instrumentation are decreasing. In the coming years, in order to make continuous monitoring efforts more feasible, research will likely focus on two major technical challenges: (1) instrument power delivery and (2) data transmission. While cabling a network via seafloor pathways is certainly one way of satisfying both requirements, cabling may not be the best long-term solution for a number of reasons. First, the costs of large cable lengths and laying them on the seafloor is often measured in the millions of dollars. Second, especially for submarine landslides, the seafloor across which the cable need be lain can be very rough and cable failure can be a quite common occurrence. Third, if cable runs are long then annual maintenance efforts and cost can also be quite high. Accordingly, techniques such as acoustic data transmission [125] and local power generation via a buoy, for instance, will be critical.

Acknowledgments

We thank the Hawaiian Volcano Observatory of the US Geological Survey, in particular Asta Miklius, Maurice Sako, and Kevan Kamibayashi, for their collaboration and efforts in operating the CGPS network at Kilauea. We thank Paul Segall of Stanford University for sharing GPS data from this network. We thank Matt Gould and Nick Witzell of WHOI for engineering support. We thank Capt. Rick Myer and the crew of the R/V Kilo Moana for their help in deploying and retrieving the extensometers. We thank Mark Schenewerk for help with implementing PAGES. We thank Katie Phillips and Dave Chadwell for helpful discussion and sharing a manuscript in preparation. We thank Cecily Wolfe and Gerard Fryer for stimulating discussion. In particular we thank J. Moore, T. Lowry, and D. Chadwell for thorough and insightful reviews which improved this contribution. This research was supported by the Geophysics and Geochemistry programs of the US National Science Foundation, grants from School of Ocean and Earth Science and Technology at the University of Hawaii, and a grant from the Deep Ocean Exploration Institute of the Woods Hole Oceanographic Institution.

Bibliography

1. Dragert H, Wang K, James TS (2001) A silent slip event on the deeper Cascadia subduction interface. Science 292:1525–1528
2. Miller MM, Melbourne TI, Johnson DJ, Sumner WQ (2002) Periodic slow earthquakes from the Cascadia subduction zone. Science 295:2423
3. Ozakawa S, Murakami M, Tada T (2001) Time-dependent inversion study of the slow thrust event in the Nankai trough subduction zone, southwestern Japan. J Geophys Res 106:782–802
4. Brooks BA, Foster JH, Bevis MF, Frazer LN, Wolfe CJ, Behn M (2006) Periodic slow earthquakes on the flank of Kilauea volcano, Hawa'i Earth and Planet Sci Lett 246:207–216
5. Cervelli P, Segall P, Johnson K, Lisowski M, Miklius A (2002) Sudden aseismic fault slip on the south flank of Kilauea volcano. Nature 415:1014–1018
6. Segall P, Desmarais EK, Shelly D, Miklius A, Cervelli P (2006) Earthquakes Triggered by Silent Slip Events on Kilauea Volcano, Hawaii. Nature 442:71–74
7. Spiess FN et al (1998) Precise GPS/Acoustic positioning of seafloor reference points for tectonic studies. Phys Earth Planet Inter 108:101–112
8. Moore JG et al (1989) Prodigious Submarine Landslides on the Hawaiian Ridge. J Geophys Res 94:17465–17484
9. Krastel S et al (2001) Submarine landslides around the Canary Islands. J Geophys Res 106:3977–3997
10. Carracedo JC, Day SJ, Guillou H, Perez Torrado FJ (1999) Giant Quaternary landslides in the evolution of La Palma and El Hierro, Canary Islands. J Volcan Geotherm Res 94:169–190
11. Day SJ, Heleno da Silva SIN, Fonseca JFBD (1999) A past giant lateral collapse and present-day flank instability of Fogo, Cape Verde Islands. J Volcan Geotherm Res 94:191–218
12. Tucholke B (1992) Massive submarine rockslide in the rift-valley wall of the Mid-Atlantic Ridge. Geology 20:129–132
13. Driscoll N, Weissel JK, Goff JA (2000) Potential for large-scale submarine slope failure and tsunami generation along the US Mid-Atlantic coast. Geology 28:407–410
14. Moore JG (1964) Giant submarine landslides on the Hawaiian Ridge. US Geol Survey Prof Paper D 501:95–98
15. Moore JG, Fiske RS (1969) Volcanic substructure inferred from dredge samples and ocean-bottom photographs, Hawaii. Geol Soc Am Bull 80:1191–1202
16. Marti J, Hurlimann M, Ablay G, Gudmundsson A (1997) Vertical and lateral collapses on Tenerife (Canary Islands) and other volcanic ocean islands. Geology 25:879–882
17. Ward SN, Day S (2001) Cumbre Vieja Volcano – Potential collapse and tsunami at La Palma, Canary Islands. Geophys Res Lett 28:3397–3400
18. Clague DA, Moore JG (2002) The proximal part of the giant submarine Wailau landslide, Molokai, Hawaii. J Volcan Geotherm Res 113:259–287
19. Coombs ML, Clague DA, Moore GF, Cousens BL (2004) Growth and collapse of Waianae Volcano, Hawaii, as revealed by exploration of its submarine flanks. Geochem Geophys Geosyst 5: doi:10.1029/2004GC000717
20. Morgan JK, Clague DA, Borchers DC, Davis AS, Milliken KL (2007) Mauna Loa's submarine western flank: Landsliding, deep volcanic spreading, and hydrothermal alteration. Geochem Geophys Geosyst 8:Q05002; doi:10.1029/2006GC001420
21. Morgan JK, Moore GF, Clague DA (2003) Slope failure and volcanic spreading along the submarine south flank of Kilauea volcano, Hawaii. J Geophys Res 108:2415, doi:10.1029/2003JB002411
22. Dieterich J (1988) Growth and Persistence of Hawaiian Volcanic Rift Zones. J Geophys Res 93:4258–4270
23. Iverson RM (1995) Can magma-injection and groundwater forces cause massive landslides on Hawaiian volcanoes? J Volcan Geotherm Res 66:295–308
24. Cannon EC, Bürgmann R, Owen SE (2001) Shallow normal faulting and block rotation associated with the 1975 Kalapana earthquake, Kilauea volcano, Hawaii. Bull Seismol Soc Am 91:1553–1562
25. Moore JG, Normark WR, Holcomb RT (1994) Giant Hawaiian landslides. Ann Rev Earth Planet Sci 22:119–144
26. Bardet J-P, Synolakis CE, Davies HL, Imamura F, Okal EA (2003) Landslide Tsunamis: Recent Findings and Research Directions. Pure Appl Geophys 160:1793–1809
27. Hasegawa HS, Kanamori H (1987) Source Mechanism of the Magnitude 7.2 Grand Banks Earthquake of November 18, 1929: Double-couple or Submarine Landslide? Bull Seismol Soc Am 77:1984–2004
28. Yeh H et al (1993) The Flores Island Tsunami. Eos, Transactions, Am Geophys Union 74:371–373
29. Imamura F, Synolakis CE, Titov V, Lee S (1995) Field Survey of the 1994 Mindoro Island, Phillipines Tsunami. Pure Appl Geophys 144:875–890
30. Geist EL (2000) Origin of the 17 July 1998 Papua New Guinea Tsunami: Earthquake or Landslide? Seismol Res Lett 71:344–351
31. Yalciner AC et al (1999) Field Survey of the 1999 Izmit Tsunami and Modeling Effort of New Tsunami Generation Mechanism. Eos, Transactions, Am Geophys Union F751(abstract):80
32. Caminade JP et al (2001) Vanuatu Earthquake and Tsunami Caused Much Damage, Few Casualties. Eos Trans Am Geophys Union 81:641,646–647
33. Nettles M, Ekstrom G (2004) Long-Period Source Characteristics of the 1975 Kalapana, Hawaii, Earthquake. Bull Seismol Soc Am 94:422–429
34. Tilling RI et al (1976) Earthquake and Related Catastrophic Events, Island of Hawaii, November 29, 1975: A preliminary report. US Geol Surv Circ 740:33
35. Day SJ, Watts P, Grilli ST, Kirby JT (2004) Mechanical models of the 1975 Kalapana, Hawaii earthquake and tsunami. Mar Geol 215:59–92, doi:10.1016/j.margeo.2004.11.008
36. Okal EA, Synolakis CE (2003) A Theoretical Comparison of Tsunamis from Dislocations and Landslides. Pure Appl Geophys 160:2177–2188
37. Ward S (2001) Landslide tsunami. J Geophys Res 106:11201–11215
38. Murty TS (2003) Tsunami wave height dependence on landslide volume. Pure Appl Geophys 160(10–11):2147–2153

39. Fine IV, Rabinovich AB, Bornhold BD, Thomson RE, Kulikov EA (2005) The Grand Banks landslide-generated tsunami of November 18, 1929, preliminary analysis and numerical modeling. Mar Geol 215:45–57
40. Phillips KA, Chadwell CD, Hildebrand JA (2008) Vertical deformation measurements on the submerged south flank of Kilauea volcano, Hawai'i reveal seafloor motion associated with volcanic collapse. J Geophys Res 113:B05106; doi:10.1029/2007JB005124
41. Owen S et al (2000) Rapid deformation of Kilauea Volcano: Global Positioning System measurements between 1990 and 1996. J Geophys Res B: Solid Earth 105:18983–18998
42. Delaney PT et al (1998) Volcanic spreading at Kilauea, 1976–1996. J Geophys Res 103:18003–18023
43. Kanamori H, Stewart GS (1979) A slow earthquake. Phys Earth Planet Inter 18:167–175
44. Sacks IS, Suyehiro S, Linde AT, Snoke JA (1978) Slow earthquakes and stress redistribution. Nature 275:599–602
45. Ihmle PF, Jordan TH (1994) Teleseismic search for slow precursors to large earthquakes. Science 266:1547–1551
46. Kedar S, Watada S, Tanimoto T (1994) The 1989 Macquarie Ridge earthquake: Seismic moment estimation from long-period free oscillations. J Geophys Res 99:17893–17908
47. McGuire JJ, Ihmle PF, Jordan TH (1996) Time-domain observations of a slow precursor to the 1994 Romanche transform earthquake. Science 274:82–85
48. Kanamori H, Cipar JJ (1974) Focal process of the great Chilean earthquake May 22:1960. Phys Earth Planet Inter 9:128–136
49. Burgmann R et al (2001) Rapid aseismic moment release following the 5 December, 1997 Kronotsky, Kamchatka, earhquake. Geophys Res Lett 28:1331–1334
50. Heki K, Miyazaki S-I, Tsuji H (1997) Silent fault slip following an interplate thrust earthquake at the Japan Trench. Nature 386:595–598
51. Segall P, Burgmann R, Matthews M (2000) Time-dependent triggered afterslip following the 1989 Loma Prieta earthquake. J Geophys Res B: Solid Earth 105:5615–5634
52. Linde AT, Gladwin MT, Johnston MJS, Gwyther RL, Bilham RG (1996) A slow earthquake sequence on the San Andreas Fault. Nature (London) 383:65–68
53. Gwyther RL, Gladwin MT, Mee M, Hart RHG (1996) Anomalous shear strain at Parkfield during 1993–94. Geophys Res Letters 23:2425–2428
54. Gao SS, Silver PG, Linde AT (2000) Analysis of deformation data at Parkfield, California: Detection of a long-term strain transient. J Geophys Res 105:2955–2967
55. Lienkaemper JL, Galehouse JS, Simpson RW (1997) Creep Response of the Hayward Fault to Stress Changes Caused by the Loma Prieta Earthquake. Science 276:2014–2016
56. Kanamori H, Kikuchi M (1993) The 1992 Nicaragua earthquake: a slow tsunami earthquake associate with subducted sediments. Nature 361:714–716
57. Lowry AR, Larson KM, Kostoglodov V, Bilham R (2001) Transient fault slip in Guerrero, southern Mexico. Geophys Res Lett 28:3753–3756
58. Lowry AR (2006) Resonant slow fault slip in subduction zones forced by climatic load stress. Nature 442(7104):802–805
59. Ito Y, Obara K, Shiomi K, Sekine S, Hirose H (2006) Slow earthquakes coincident with episodic tremors and slow slip events. Science 26:503–506
60. Rogers G, Dragert H (2003) Episodic Tremor and slip on the Cascadia Subduction Zone: The chatter of silent slip. Science 300:1942–1943
61. Ide S, Beroza GC, Shelly DR, Uchide T (2007) A scaling law for slow earthquakes. Nature 447:76–79; doi:10.1038/nature05780
62. Aki K, Fehler M, Das S (1977) Source mechanism of volcanic tremors: fluid driven crack models and their application to the 1963 Kilauea eruption. J Volcan Geotherm Res 141:259–287
63. Rubenstein JL et al (2007) Non-volcanic tremor driven by large transient shear stresses. Nature 448:579–582
64. Shelly DR, Beroza GC, Ide S (2007)Non-volcanic tremor and low-frequency earthquake swarms. Nature 446:305–307; doi:10.1038/nature05666
65. Peacock S, Wang K (1999) Seismic Consequences of Warm Versus Cool Subduction Metamorphism: Examples from Southwest and Northeast Japan. Science 286:937–939
66. Julian B (2002) Seismological detection of slab metamorphism. Science 296:1625–1626
67. Kato N (2004) Interaction of slip on asperities: Numerical simulation of seismic cycles on a two-dimensional planar fault with nonuniform frictional property. J Geophys Res 109:B12306; doi:10.1029/2004JB003001
68. Liu Y, Rice JR (2005) Aseismic slip transients emerge spontaneously in three-dimensional rate and state modeling of subduction earthquake sequences. J Geophys Res 110:B08307; doi:10.1029/2004JB003424
69. Miyazaki S, McGuire JJ, Segall P (2003) A transient subduction zone slip episode in southwest Japan observed by the nationwide GPS array. J Geophys Res 108(B2):2087; doi:10.1029/2001JB000456
70. Kostoglodov V et al (2003) A large silent earthquake in the Guerrero seismic gap, Mexico. Geophys Res Lett 30:1807, doi:10.1029/2003GL017219
71. Freymueller J et al (2001) The great alaska 'earthquake' of 1998–2001. EOS Trans Am Geophys Un 82(47)
72. Hirose H, Hirahara K, Kimata F, Fujii F, Miyazaki S (1999) A slow thrust slip event following the two 1996 Hyuganada earthquakes beneath the Bungo Channel, southwest Japan. Geophys Res Lett 26:3237–3240
73. Baum RL, Reid ME (1995) Geology, hydrology, and mechanics of a slow-moving, clay-rich landslide, Honolulu, Hawaii. In: Haneberg WC, Anderson SA (eds) Clay and Shale Slope Instability. Geol. Soc. of Am, Boulder, Colorado
74. Malet JP, Maquaire O, Calais E (2002) The use of Global Positioning System techniques for the continuous monitoring of landslides: application to the Super-Sauze earthflow (Alpes-de-Haute-Provence, France). Geomorphology 43:33–54
75. Hilley GE et al (2004) Dynamics of slow-moving landslides from permanent scatterer analysis. Science 304(5679):1952–1955
76. Chadwell CD et al (1999) No spreading across the southern Juan de Fuca Ridge axial cleft during 1994–1996. Geophys Res Lett 26:2525–2528
77. Chadwick WW, Embley RW, Milburn HR, Meining C, Stapp M (1999) Evidence for deformation associated with the 1998 eruption of Axial Volcano, Juan de Fuca Ridge, from acoustic extensometer measurements. Geophys Res Lett 26:3441–3444

78. Gagnon K, Chadwell CD, Norabuena E (2005) Measuring the onset of locking in the Peru-Chile trench with GPS and acoustic measurements. Nature 434:205–208
79. Sweeney AD, Chadwell CD, Hildebrand JA, Spiess FN (2005) Centimeter-level positioning of seafloor acoustic transponders from a deeply-towed interrogator. Mar Geol 28:39–70, doi:10.1080/01490410590884502
80. Cervelli P, Murray MH, Segall P, Aoki Y, Kato T (2001) Estimating source parameters from deformation data, with an application to the March 1997 earthquake swarm off the Izu Peninsula, Japan J Geophys Res 106:11217–11237
81. Dixon TH (1991) An Introduction to the Global Positioning System and Some Geological Applications. Rev Geophys 29:249–276
82. Hager BH, King RW, Murray MH (1991) Measurement of Crustal Deformation Using the Global Positioning System. Ann Rev Earth Planet Sci 19:351–382
83. Segall P, Davis JL (1997) GPS applications for geodynamics and earthquake studies. Ann Rev Earth Planet Sci 25:301–336
84. Eckl MC, Snay RA, Soler T, Cline MW, Mader GL (2001) Accuracy of GPS-derived relative positions as a function of interstation distance and observing-session duration. J Geodesy 75:633–640
85. Leick A (2004) GPS satellite surveying. Wiley, Hoboken
86. Warnant R, Kutiev I, Marinov P, Bavier M, Lejeune S (2007) Ionospheric and geomagnetic conditions during periods of degraded GPS position accuracy: 1. Monitoring variability in TEC which degrades the accuarcy of Real-Time Kinematic GPS applications. Adv Space Res 39:875–880
87. Bock Y, Nikolaidis RM, de Jonge PJ, Bevis M (2000) Instantaneous geodetic positioning at medium distances with the Global Positioning System. J Geophys Res 105:28233–28253
88. Langbein, J (2004) Noise in two-color electronic distance meter measurements revisited. J Geophys Res 109:B04406; doi:10.1029/2003JB002819
89. Brooks BA, Foster JF, Miklius A, Schenewerk M (2005) Extended GPS Network and Near Real-Time processing, Mauna Loa Volcano, Hawai'i. Eos Trans AGU, 86:Fall Meet Suppl
90. Schenewerk M, Dillinger W, Hilla S (2000) http://www.ngs.noaa.gov/GRD/GPS/DOC/toc.html
91. Mader G, MacKay JR (1996) Geoscience Laboratory, Office of Ocean and Earth Sciences NOS, NOAA. Silver Spring, Maryland
92. McCarthy D, Petit G (eds) (2003) IERS Technical Note 32
93. Cartwright DE, Edden AC (1973) Corrected tables of tidal harmonics. Geophys J R Astron Soc 33:253–264
94. Cartwright DE, Taylor RJ (1971) New computation in the tide-generating potential. Geophys J R Astron Soc 23:45–74
95. Foster J, Bevis M, Chen Y-L, Businger S, Zhang Y (2003) The Ka'u Storm (Nov 2000): Imaging precipitable water using GPS. J Geophys Res 108:4585
96. Segall P, Matthews M (1997) Time dependent inversion of geodetic data. J Geohys Res 102:22391–22409
97. Flores A, Ruffini G, Rius A (2000) 4D tropospheric tomography using GPS slant wet delays. Ann Geophys 18:223–234
98. Spiess FN (1985) Analysis of a possible sea floor strain measurement system. Mar Geodesy 9:385–398
99. Zumberge M (1997) Precise Optical Path Length Measurement Through an Optical Fiber: Application to Seafloor Strain Monitoring. Ocean Eng 24:532–542
100. Zumberge M et al (2006) In: Nadim F, Pottler R, Einstein H, Klapperich H, Kramer S (eds) 2006 ECI Conference on Geohazards. Lillehammer, Norway
101. Okada Y (1985) Surface deformation due to shear and tensile faults in a half-space. Bull Seismol Soc Am 75:1135–1154
102. Steketee JA (1958) Some Geophysical Applications of the Elasticity Theory of Dislocations. Can J Phys 36:1168–1198
103. Menke W (1984) Geophysical Data Analysis: Discrete Inverse Theory. Academic Press Inc, San Diego
104. Brooks BA, Frazer LN (2005) Importance reweighting reduces dependence on temperature in Gibbs samplers: an application to the coseismic geodetic inverse problem. Geophys J Int 161:12–21
105. Mosegaard K, Sambridge M (2002) Monte Carlo analysis of inverse problems. Inverse Probl 18:29–54
106. Kirkpatrick S, Gelatt CDJ, Vecchi MP (1983) Optimization by simulated annealing. Science 220:671–680
107. Hudnut KW et al (1996) Co-Seismic Displacements of the 1994 Northridge, California, Earthquake. Bull Seismol Soc Am 86:19–36
108. Dieterich J, Cayol V, Okubo PG (2000) The use of earthquake rate changes as a stress meter at Kilauea volcano. Nature 408:457–460
109. Stein RS (1999) The role of stress transfer in earthquake occurrence. Nature 402:605–609
110. Hansen S, Thurber CH, Mandernach MJ, Haslinger F, Doran C (2004) Seismic velocity and attenuation structure of the East Rift Zone and South Flank of Kilauea Volcano, Hawaii. Bull Seismol Soc Am 94:1430–1440
111. Got J-L, Okubo PG (2003) New insights into Kilauea's volcano dynamics brought by large-scale relative relocation of microearthquakes. J Geophys Res 108(B7):2337; doi:10.1029/2002JB002060
112. Wolfe CJ, Brooks BA, Foster JH, Okubo PG (2007) Microearthquake streaks and seismicity triggered by slow earthquakes on the mobile south flank of Kilauea Volcano, Hawai'i. Geophys Res Lett (In press)
113. McGuire JJ, Segall P (2003) Imaging of aseismic fault slip transients recorded by dense geodetic networks. Geophys J Int 155:778–788
114. Scholz CJ (1998) Earthquakes and friction laws. Nature 391:37–42
115. Perfettini H, Schmittbuhl J, Rice JR, Cocco M (2001) Frictional response induced by time-dependent fluctuations of the normal loading. J Geophys Res 106:435–438
116. Liu Y, Rice JR (2007) Spontaneous and triggered aseismic deformation transients in a subduction fault model. J Geophys Res 112:B09404; doi:10.1029/2007JB004930
117. Rice JR, Gu JC (1983) Earthquake aftereffects and triggered seismic phenomena. Pure Appl Geophys 121:187–219
118. Perfettini H, Schmittbuhl J (2001) Periodic loading on a creeping fault: Implications for tides. Geophys Res Lett 28:435–438
119. Shibazaki B, Shimamoto T (2007) Modelling of short-interval silent slip events in deeper subduction interfaces considering the frictional properties at the unstable-stable transition regime. Geophys J Int 171(1):191–205
120. Elsworth D, Day SJ (1999) Flank collapse triggered by intrusion: the Canarian and Cape Verde Archipelagoes. J Volcan Geotherm Res 94:323–340

121. Cervelli P et al (2002) The 12 September 1999 upper east rift zone dike intrusion at Kilauea Volcano, Hawaii. J Geophys Res B: Solid Earth 107:3-1–3-13

122. Owen S et al (2000) January 30, 1997 eruptive event on Kilauea Volcano, Hawaii, as monitored by continuous GPS. Geophys Res Lett 27:2757–2760

123. McMurtry GM, Watts P, Fryer GJ, Smith JR, Imamura F (2004) Giant landslides, mega-tsunamis, and paleo-sea level in the Hawaiian Islands. Mar Geol 203:219–233

124. Ward S (2002) Slip-sliding away. Nature 415:973–974

125. Frye D et al (2006) An Acoustically Linked Moored–Buoy Ocean Observatory. EOS Trans AGU 87:213–218

Swarm Intelligence

Gerardo Beni
University of California Riverside, Riverside, USA

Article Outline

Glossary

Swarm Intelligence (Definition 1, Sect. "Definition of the Subject"). The *intuitive* notion of "Swarm Intelligence" is that of *a "swarm" of agents (biological or artificial) which, without central control, collectively (and only collectively) carry out (unknowingly, and in a somewhat-random way) tasks normally requiring some form of "intelligence"*. (Definition 5, Sect. "Swarms of Intelligent Units"). The capability of universal computation carried out with natural asynchrony by a dynamic cellular computing system, none of whose cells can predict the computation done by the swarm.

Swarm Robotics The technology of robotic systems capable of Swarm Intelligence.

Stigmergy Indirect communication through modification of the environment.

Pheromone A chemical that triggers an innate behavioral response in another member of the same animal species

Elementary Swarm An ordered set of N units described by the N components v_i ($i = 1, 2, \ldots, N$) of a vector $\boldsymbol{v}$; any unit i may update the vector, at any time t_i, using a function f of K_i vector components.

$$\forall i \in N\colon\ v_i(t+1) = f(\mathbf{v}_{k \in K(i)}(t))$$

Cellular automaton A system, evolving in discrete time steps, with four properties: a grid of cells, a set of possible states of the cells, a neighborhood, and a function which assigns a new state to a cell given the state of the cell and of its neighborhood.

von Neumann architecture Computer design that uses one processing unit and one storage unit holding both instructions and data.

Cellular-computing architecture Computer design that uses cellular automata, and related machines, as processors and as storage of instruction and data.

Intelligence (working definition for Swarm Intelligence) Ability to carry out universal computation.

Natural asynchrony Asynchronous updating characterized by three properties: more than one unit may update at each time step; any unit may update more than once in each updating cycle; and the updating order varies randomly for every updating cycle.

Optimization algorithms Algorithms to satisfy a set of constraints and/or optimize (e. g., minimize) a function by systematically choosing the values of the variables from an allowed set.

Swarm optimization Ant colony optimization, Particle swarm optimization, and related probabilistic optimization algorithms.

Ant colony optimization Probabilistic optimization algorithm where a colony of artificial ants cooperate in finding solutions to optimization problems.

Particle swarm optimization Probabilistic optimization algorithm where a swarm of potential solutions (particles) cooperate in finding solutions to discrete optimization problems.

Game of Life A cellular automaton designed to simulate life-like phenomena.

Dynamic cellular computing system Cellular computing system whose cells are mobile.

Unpredictable system A system such that complete knowledge of its state and operation at any given time is insufficient to compute the system's future state before the system reaches it.

Definition of the Subject

The research area identified as "Swarm Intelligence" has evolved rapidly. The term "Swarm Intelligence" first appeared in 1989 [1,2]; by 2007 "Swarm Intelligence" was in the title of four books [3,4,5,6], two series of conference proceedings [7,8] and in a new technical journal [9], without mentioning other areas in which the term swarm itself has become popular. As the use of the term "Swarm Intelligence" has increased, its meaning has broadened to a point in which it is often understood to encompass almost any type of collective behavior. And since the term "Swarm Intelligence" has popular appeal, it is also sometimes used in contexts which have limited scientific or technological content. Other meanings, however, refer rigorously to precise concepts. The following treatment of Swarm Intelligence is based only on concepts that can be clearly defined and quantified. Hence it is more restricted than some broader Swarm Intelligence presentations but, even so, it describes an interrelated scientific/technical core which forms a solid basis for a well-defined multidisciplinary research area.

Definition 1 The intuitive notion of "Swarm Intelligence" is that of a "swarm" of agents (biological or artificial) which, without central control, collectively (and only collectively) carry out (unknowingly, and in a somewhat-random way) tasks normally requiring some form of "intelligence".

A more specific definition requires a detailed discussion and so it is given at the end of the article. (Sects. "Characteristics of Swarm Intelligence" and "Swarms of Intelligent Units").

Although this notion of Swarm Intelligence might seem vague, we will see in the course of this article that in fact it has many specific implications. Note that the notion is broad, which partly explains its widespread use, but not so broad as to include any type of collective action of groups of simple entities, as will become clear later.

These characteristics of Swarm Intelligence are also those of several biological systems, e. g., some insect societies or some components of the immune system, so that Swarm Intelligence has become important for understanding certain mechanisms in biology.

Technologically, the importance of 'swarms' is mainly based on the potential advantages over centralized systems. The potential advantages are: (1) *economy*; the swarm components (units) are simple, hence, (in principle) mass producible, modularizable, interchangeable, and disposable; (2) *reliability*, due to the redundancy of the components; destruction/death of some units has negligible effect on the accomplishment of the task, as the swarm adapts to the loss of few units; (3) ability to perform *tasks beyond those of centralized systems*, e. g., escaping enemy detection.

From this initial perspective on potential advantages, the actual application of Swarm Intelligence has extended to many areas, described in the body of this article, and its potential for future applications remains high, as discussed in the concluding section.

Some current, proposed and/or potential applications are in defense and space technologies, (e. g., control of groups of unmanned vehicles in land, water, or air), flexible manufacturing systems, advanced computer technologies (biocomputing), medical technologies, and telecommunications.

Introduction

Swarm Intelligence (SI) investigations were initially motivated by studies of groups of simple robotic units which offered promise for technology. These 'swarms' are modeled as collections of simple quasi-identical units with decentralized control and independent clocks [10]. The number of units is intermediate between those of typical systems investigated in physics and other traditional science fields; in fact, the swarm is of the order of 10^2 to 10^s units, where $s \ll 23$, i. e., the swarm is not composed of so many units that statistical physics methods can be applied to it, nor of such a few units that its dynamic can be solved exactly (or numerically to high precision).

These features are typical also of many biological systems, such as insect societies. They are also the features of robotic systems potentially economic, reliable and capable of tasks beyond the capabilities of centralized systems.

There are various established fields of science and technology that deal to a certain extent with the intuitive notion of SI: Artificial Life, Ethology, Robotics, Artificial Intelligence, Computation, Self-Organization, Complexity, Economics, and Sociology. In all these fields there is at some level, or in some application, the need to understand, model, and predict the operation of groups of units which only by working together (in a not very structured way, and 'unaware' of the evolution of the group) carry out "intelligent" tasks. A simple illustration is provided by classic economics. In a free market economy, people trade with each other in an unstructured way: each makes independent decisions at unpredictable times and unaware of the global results. But the outcome is the solution to a complex problem – the problem of correct pricing. Such a 'swarm' solves a problem not (or poorly) solvable by centralized control economies, thus exhibiting, in a certain sense, a high form of "intelligence".

From this example it is clear that the intuitive notion of SI is easy to grasp. But it is not so easy to make it less vague and more precise and quantitative, i. e., to make it a useful working concept for science and technology. That the concept of SI is not easy to quantify follows from the difficulty of defining several of the key components of the intuitive notion of SI (Definition 1). First, "intelligence" is a notoriously ambiguous concept. Second, "randomly" is also not easily defined and quantified. Third, "only collectively" must be specified in terms of the critical number of agents required for the emergence of SI. In what sense is a unit 'simple' or, 'un-intelligent', and the task carried out 'complex' or, 'intelligent'? Fourth, "unknowingly" implies that the global status and the goal of the swarm are, at least to some extent, unknown to the single agents. Which algorithms and communication schemes result in tasks carried out 'unknowingly' by the agents?

Because of these difficulties, in this article we first use the aforementioned (Definition 1) intuitive notion of SI to describe the current main areas of studies considered to be SI. This will provide an overview of the current status of the field; it will make it possible to quantify the four vague concepts ('intelligence', 'randomly', 'collectively', 'unknowingly') and to reach a more sharply defined concept of SI. From this, we will be able to see more clearly the limitations of SI, and so, its realistic potential for future applications.

In this article, the main areas of SI studies are described by making three very broad distinctions: (1) scientific interest vs. technological interest (Sects. "Biological Systems"–"Definition of Swarm"); (2) standard mathematics vs. cellular computational mathematics (Sects. "Standard-Mathematics Methods"–"Cellular-Computing Methods"); (3) synchronous operation vs. asynchronous operation (Sects. "Randomness in Swarm Intelligence"– Swarms of Intelligent Units"). These distinctions in turn will provide a guide to clarifying the four vague concepts in the intuitive definition of SI (Definition 1) and, thus, a conceptual orientation for future studies and applications; they will also provide criteria for evaluating the promise of SI to solve complex problems that traditional approaches cannot.

Focusing on the first distinction (scientific vs. technological interest), the main scientific interest in SI originated with the work of biologists studying insect societies [3]. The main technological interest originated with roboticists trying to design distributed robotic systems [1,2]. A valuable reference on the development of SI is [3], dealing, in parallel, with these two interrelated interests.

Biological Systems

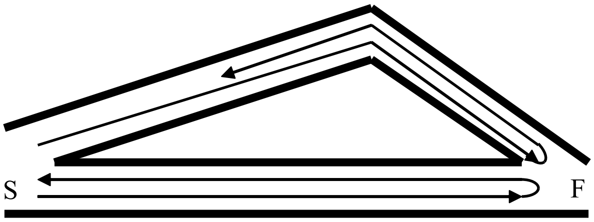
Probably the best known and seminal biology experiment in SI is the 'double bridge' experiment by Goss et al. [11]. While studying the foraging of ants they observed that if ants, starting from a point *S*, could reach food at a point *F* via two paths of different lengths, the ants would at first choose one of the two paths randomly; but after some ants had returned from *F* to *S*, more ants would choose to go from *S* to *F* via the shortest path; and eventually practically all the ants would choose the shortest path (see Fig. 1).

The key insight was the realization that the ants were finding the best path via *stigmergy*, that is, by *communicating through modification of the environment*. Ants are blind but they communicate chemically via pheromones. By laying pheromones along the path when returning from the food source F, the ants effectively marked the shortest path by laying more pheromones on it. After that, the ants that would start from *S*, would choose the path marked by more pheromones, i. e., the shortest path.

Thus it was observed and understood a method of self-organization and a method to solve a nontrivial problem by a form of collective intelligence, with many of the elements of the intuitive definition of SI given above.

Later, Dorigo [12] realized that this method could be abstracted and generalized to design algorithms that rely on 'artificial ants' to solve much more complex problems (see Sect. "Swarm Optimization"). Thus the close connection between biological studies of SI and its potential for technological application was first clearly demonstrated.

Many other experiments in ants and other social insects have confirmed the potential for developing bio-inspired algorithms [3,13,14,15]. For actual insect societies, various ant algorithms have been applied to model



Swarm Intelligence, Figure 1
Illustration of the double bridge experiments. The ants following the shorter path (the lower path) return to the source before the ants which have taken the longer path. In this way the shorter path has a higher density of pheromones; as a result, ants starting at *S* will now prefer the shorter path

tasks such as: division of labor, cemetery organization, brood care, carrying of large objects, constructing bridges, foraging, patrolling, chaining, sorting eggs, nest building and nest brooming. Social insects constitute 2% of insects, half being ants. Besides ants, termites, bees, and wasps, have been observed to exhibit some forms of SI behavior as in the aforementioned tasks.

Apart from insects, many other biological groups exhibit behavior with some of the features of SI, such as flocks of birds and schools of fish. A seminal model of artificial flocks and schools of fish was proposed by Craig Reynolds in 1987 [16].

It is a computational model for simulating the animation of a group of entities called "boids", i. e., it is intended to represent the group movement of flocks of birds and fish schools. In this model, each boid makes its own decisions on its movement according to a small number of simple rules that react to the neighboring members in the flock and the environment it can sense. The simple local rules of each boid generate complex global behaviors of the entire flock. In addition to being used to simulate group motion in a number of movies and games, this flocking behavior has been used, e. g., for time-varying data visualization [17].

In studying these biological systems several concepts of relevance to SI were recognized. They can be summarized as:

(1) Multiple communication (of various types) among units;
(2) Randomness (random fluctuations);
(3) Positive feedback, to reinforce random fluctuations;
(4) Negative feedback for stabilization.

Of the various types of communication, we have already noted *stigmergy*, i. e., *indirect communication* by modification of the environment. On the other hand, *direct communication* may occur *unit-to-unit*, contact being a special case, (e. g., via antennae or mandibles in insects) or by *broadcasting* within a certain range (e. g., acoustically or chemically). The type and specific mode of communication has been found to be critical to the task performed, as e. g., in what types of patterns are formed [18].

Finally, the most basic lesson from biological studies of SI is that biology has found solutions to hard computational problems, and that the design principles used in doing this, can be imitated.

Robotic Systems

The actual realization of SI systems as collections of robots is a very hard problem; in fact, it is quite difficult to make even small groups of robots perform useful tasks [19,20]. Making even a single mobile, autonomous robot work in a reliable way (even in simplified environments) is a complex project. Often the technical problems with small groups of robots are quite far from the goal of SI, so there is not much reason to use the term 'swarm'. Terms such as "collective robotics", "multi-robot systems", and "distributed autonomous robotic systems" are generally, and more appropriately, used. But, whenever the tasks carried out by these robotic systems become scalable to large numbers, the term "swarm robotics" is appropriate and, in fact, it has come into use. More typically, "swarm robotics" simply describes the design of groups of robotic units performing a collective task. Each robotic unit cannot solve the task alone; and collectively the robotic units try to accomplish a common task without centralized control.

As for any robotic system in general, each robotic unit, and the group as a whole, require design of: mechanics, control, and communications. The emphasis of current research, in relation to swarm robotics, is primarily on the latter two: (1) effective *communication* among the robot units [21], and (2) effective *control* via decentralized algorithms and robustness [22].

(1) Research in robotic *communication* has become important with the growth of wireless communication networking and the lower cost of building robotic units, thus opening a new range of applications for multi-robot systems with networking capabilities, including swarm robotics. In fact swarm robotics provides the common ground for convergence of information processing, communication theory and control theory [21].
(2) Research in *control* of robotic swarms is particularly important to guarantee the stability of the swarm since the swarm does not have a centralized control. The stability of a swarm is a special case of the general problem of distributed control. In fact, after swarm robotics algorithms for task implementation have been devised, the practical realization requires stability, and robustness, i. e., proper control. Swarm control presents new challenges to robotics and control engineers: various types of controllers for swarms are currently being investigated, e. g. neural controllers [23].

The control theory example coming closest to the problem of swarm control is perhaps that of 'formation' control, e. g., the control of multi-robot teams or autonomous aircrafts or land or water vehicles. These studies, when extended to decentralized systems, lead to consider problems of asynchronous stability of distributed robotic systems and swarms. [24]

Although much progress has been made in swarm robotics, the application of SI algorithms is still underdeveloped; one reason is that often the SI behavior emerges only above a critical number which is too large to make the construction of a robotic swarm practical, because too complex or expensive. Investigations of this type are thus generally carried out by simulation [22,25].

These simulations are specialized methods of swarm robotics. An example is 'executable models' [25] which can run in simulation or on a mobile robotic unit and can execute all aspects of a robotic unit behavior (sensing, information processing, actuation, motion), i. e., they fully represent how perception is translated into action by each robotic unit. They can be used to test either experimentally or in simulation how a group of robotic units behaves. Executable models are an evolution of early protocols (so called 'behavior-based' protocols) designed around the subsumption architecture [26]. Behavior-based protocols have now been generalized into Markov-type methods, i. e., protocols where the transitions between the possible states of a robotic unit are specified by a probability transition matrix as in Markov processes [27].

Looking at applications, swarm robotics has by now accumulated a collection of standard problems which recur often in the literature. One group of problems is based on *pattern formation*: aggregation, self-organization into a lattice, deployment of distributed antennas or distributed arrays of sensors, covering of areas, mapping of the environment, deployment of maps, creation of gradients etc. A second group of problems focuses on *some specific entity in the environment*: finding the source of a chemical plume, homing, goal searching, foraging, prey retrieval, etc. And a third group of problems deals with more *complex group behavior*: cooperative transport, mining (stick picking), shepherding, flocking, containment of oil spills, etc. This is not an exhaustive list: other generic robotic tasks, such as obstacle avoidance and all terrain navigation, are also swarm robotics tasks.

One envisioned application of swarm robotics which has received media attention is the ANTS (autonomic nanotechnology swarm) project by NASA [28,29]. This project envisions *nanobots* (i. e., a swarm of microscopic robots) operating autonomously to form structures for space exploration. The idea is inspired by the example of insect societies; it envisions a technology of self-similar, reconfigurable, miniaturized robotic units with a software strategy to endow the swarm with 'intelligence'. This ANTS 'intelligence' is at an intermediate level between traditional Artificial Intelligence (i. e., highly symbolic) and reactive responses 'intelligence', i. e., intelligence without internal representation [30]. Its basis is a software construct called a 'neural basis function' to bridge the gap between lower and higher level functions and to be capable of autonomous behavior. In one potential implementation, a Saturn autonomous ring array would launch 1,000 spacecraft with specialized instruments—organized as 10 subswarms—to perform in situ exploration of Saturn's rings to understand their constitution and formation.

The European Union sponsored swarm robotics project [31,32] was completed in 2005 after demonstrating several critical tasks, such as: autonomous self-assembly, cooperative obstacle avoidance, and group transport. For this project a new type of robot, called an s-bot was developed. A swarm-bot is any device composed of more than one s-bot, which is a mobile robot unit capable of connecting or disconnecting from another s-bot. S-bots have relatively simple sensors and motors and limited computational capabilities. Using their grippers, s-bots can assemble into a swarm-bot that is able to solve problems too difficult for a single s-bot. For example, a swarm-bot could transport an object too heavy for a single s-bot. [33]

Although swarm robotics could be defined as the robotic implementation of SI (Definition 1), so far, as noted, this implementation remains a distant goal. Meanwhile, concepts from SI can be usefully applied to collections of cooperating robots. Thus, referring to the intuitive notion of SI (Definition 1), the robotic swarm, can be characterized by the type of algorithm and of (decentralized) control, the number of units above which new behavior emerges, the communication method (range, topology, bandwidth), the processing and memory capability of each unit, and the heterogeneity of the group.

Swarm robotics, besides the implementation of SI algorithms, includes the material (mechanical and electronic) realization of the units comprising the swarm. This is, as noted, an arduous task which often becomes the emphasis of research in swarm robotics. But, as it was emphasized in the early years of SI, even if the material construction of the swarm were accomplished, SI algorithms would remain as the most difficult challenge for swarm robotics. This can be easily seen from the fact that a 'robot' swarm with very advanced hardware is already available for experimentation: it is a group of human beings. Each person could be limited in a controlled way, e. g., by allowing each person to handle only a specific device according to specific rules. Algorithms to make such a swarm doing intelligent tasks are in the province of SI, but they are not simple to devise, as common experience shows.

Artificial Life Systems

The areas of *Self-organization*, *Complexity* and *Artificial Life (or A-life)* are all older and broader fields than SI and overlap with it to various extents.

A-life is conceptually placed somewhere between science and technology, and between biology and robotics. During the mid-1980s attempts at imitating living systems with machines grew rapidly and resulted in the formation of the research field of "Artificial Life" [34]. A-life investigates phenomena characteristic of living systems primarily through computational and (to a lesser extent) robotic methods.

Its scope is wide, ranging from investigations of how life-like properties develop from inorganic components to how cognitive processes emerge in natural or artificial systems. It includes research on any man-made systems that mimic the characteristics of natural living systems. By this criterion it includes SI, but actual, current A-life research is not much focused on SI; rather it focuses on origin and synthesis of life; evolutionary robotics; morphogenesis, learning, etc.

The basic theories at the foundation of A-life, and of relevance to SI, are the theories of *self-organization* and *complexity*. A-life studies systems which are typically characterized by many strongly coupled degrees of freedom. Systems of this type are more generally investigated within the science of *complexity* which began to be an active field of research in the early '80s. It is multidisciplinary and it investigates physical, biological, computational, and social science problems, including a vast range of topics [35] from environmental sciences to economics as it is clear from the content of this Encyclopedia. One basic feature that these systems have in common is the emergence of complex behavior from simple components, a notion we also find in SI.

In regard to self-organization, we note that, as many systems in nature, A-life systems may start disordered and featureless, and then spontaneously organize themselves to produce ordered structures, i. e., they self-organize. The theory of self-organization, going back to the 1950's [36], grew out of a variety of disciplines, but mainly from thermodynamics, non-linear dynamics and control theory. Self-organization can be defined as the spontaneous creation of a globally coherent (i. e., entropy lowering) pattern out of local interactions – a concept also relevant to SI.

Because of its distributed character, self-organization tends to be robust, resisting perturbations. The dynamics of a self-organizing system is typically non-linear, because of feedback relations between the components. Positive feedback leads to fast growth, which ends when all components have been absorbed into the new configuration, leaving the system in a stable, negative feedback state. Non-linear systems have in general several stable states, and this number tends to increase (bifurcate) as an increasing input of energy forces the system away from its thermodynamic equilibrium. To adapt to a changing environment, the system needs a variety of stable states that is large enough to react to perturbations but not so large as to make its evolution uncontrollably chaotic. The most adequate states are selected according to their fitness, either directly by the environment, or by subsystems that have adapted to the environment at an earlier stage.

Formally, the basic mechanism underlying self-organization is the (often driven by randomness) variation which explores different regions in the system's state space until it enters an attractor. This precludes further variation outside the attractor, and thus restricts the freedom of the system's components to behave independently. It is equivalent to the decrease of statistical entropy that defines self-organization.

It is useful to keep this brief sketch of self-organization theory in mind as we proceed in describing SI, since the concepts in the theory of SI are evolved from a combination of concepts of self-organization and computation.

Definition of Swarm

After having looked, in the previous three sections, at actual robotic, biological, and A-life systems and ideas of complexity and self-organization related to SI, we can return to the intuitive definition of SI (Definition 1) and make it more quantitative.

The intuitive notion consists of four elements: SI is "intelligence" achieved "collectively", "randomly", and "unknowingly". An elementary swarm retaining these four elements can be defined as

Definition 2 Elementary Swarm An ordered set of N units described by the N components v_i $(i = 1, 2, \dots, N)$ of a vector $\mathbf{v}$; any unit i may update the vector, at any time t_i, using a function f of K_i vector components. $\forall i \in N\colon \mathbf{v}_i(t+1) = f(v_{k\in K(i)}(t))$

The Elementary Swarm describes an internally driven "collective" action. External input may be added in f. "Randomness" is built in the updating times. The evolution occurs "unknowingly" since the units have no processing capability. The Elementary Swarm can be generalized so that randomness appears also in the parameters of the function f. A further generalization is obtained by letting each vector component to be not just one number but a set of parameters.

Hereinafter we call '*Swarm*' (capital S) any system capable of SI. It is worth noting that even in Swarms more general than the Elementary Swarm, the modeling is assumed restricted in such a way that *no unit is capable of computing the Swarm's next global state,* (see also Sect. "Swarms of Intelligent Units"). Finally, "intelligence" is expected to be achieved by running appropriate algorithms via the updating function f. If and how this is going to be possible requires a more mathematical discussion, which is the subject of Sects. "Standard-Mathematics Methods"–"Cellular-Computing Methods".

Standard-Mathematics Methods

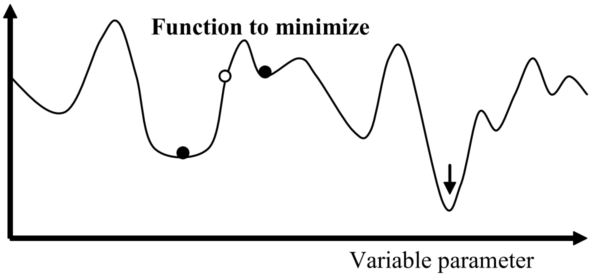
The science of biological swarms and the engineering of robotic swarms, as well as research in A-life relevant to SI, have progressed by using a broad range of mathematical techniques. All these techniques can be classified in two main groups: (1) '*standard-mathematics' methods* and (2) *cellular-computational methods.*

By *standard-mathematics methods* (SMm) we mean any method that is based on the standard tools of applied mathematics and computations based on standard (Von Neumann) computer architectures. Examples are methods in differential equations, stochastic techniques, linear systems, and optimization. By *cellular-computing methods* (CCm) we mean highly parallel and local computational methods, with simple cells as the basic units of computation, typically carried out on ▶ Mathematical Basis of Cellular Automata, Introduction to (CA) [37].

These two mathematical approaches reflect two distinct trends in the evolution of SI research, as described below. We consider first (Sects. "Swarm Optimization"–"Limitations of Standard-Mathematics Methods") the approach to SI based on SMm, since the greatest number of significant results in the area of SI has been obtained, so far, by standard-mathematics methods; specifically in the areas of *optimization* and *non-linear dynamics*. We consider them in turn in the next two sections.

Swarm Optimization

Optimization is by far the largest research area associated with SI. This is due to two extremely successful optimization methods, whose origin is related to models of SI. The two methods are the Ant Colony Optimization (ACO) [12,13] and the Particle Swarm Optimization (PSO) [4,38]. Both ACO and PSO, originated in the early nineties and have resulted in hundreds of applications based on variations of the original algorithms. So much so that the field of "Swarm Optimization" could stand alone, apart from its relation to SI with which it is sometimes even identified. A thorough and recent description of swarm optimization techniques is in [6]. Here only the key concepts of swarm optimization are reviewed, as they relate to SI.



Swarm Intelligence, Figure 2
Simplified illustration of the typical problem encountered in optimization. Starting from the value represented by the *open circle* and varying the parameter continuously the algorithm will find one of the nearest local minima (*black circles*) rather than the global minimum (indicated by the *arrow*)

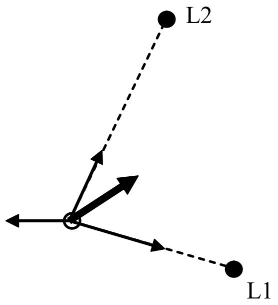
In PSO and ACO, as in any optimization methods, a function must be optimized, e. g., minimized. To find the minimum of the function, the variable is changed in a systematic way – the optimization method. Generally, the variable spans a multidimensional space. The search for the global minimum is nontrivial since the function may have many local minima, and the search could end into one of them (see Fig. 2). Various techniques to avoid this trapping have been developed by using some degree of randomness in the search strategy. For example, simulated annealing [39] was developed to overcome the limitations of non-random methods, e. g., the gradient descent [40]. PSO and ACO belong to this class of optimization techniques that make use of randomized searches.

Particle Swarm Optimization (PSO)

PSO was developed by Kennedy and Eberhart in 1995 [38] inspired by the social behavior of bird flocking and fish schooling [16].

In PSO, the position of each unit of the swarm is a point in the variable space of the function to be minimized. Every unit tries to reach the position corresponding to the minimum of the function. Each unit is assumed to know the global minimum value of the function, and to detect the value of the function at its location as well as the value of the function at the locations of a group of neighbors. The size of the group of neighbors is a parameter and could be the whole swarm.

The algorithm is, schematically, as follows. The units are initially in random locations. Every unit moves in



Swarm Intelligence, Figure 3
Illustration of the velocity update mechanism in PSO. A unit (*white circle*) originally moving to the left, changes its velocity by adding the two components in the direction of L1 (the unit's best location so far) and of L2 (the neighbors' best location so far). The magnitude of these components is determined by random weights. The new velocity (*heavy solid line*) generally tends to be in the direction of the global optimum

the variable space and remembers the location L1 where, among the locations visited so far, the function had minimum value. It also remembers the location L2 where, among the locations visited by all its neighbors, the function had minimum value.

At each time step, each unit calculates its distances from L1 and from L2; forms a weighted average of L1 and L2 using random weights; and changes its velocity in proportion to this weighted average (see Fig. 3). As a result every unit tends to move toward the location of the global minimum by taking advantage of its and its neighbors' knowledge. The process stops either when a unit is sufficiently close (by a chosen tolerance) to the location of the global minimum or when a chosen maximum number of iterations has been run.

PSO belongs to the category of stochastic, population-based algorithms, such as, e. g., genetic algorithms. The PSO's great merit is its simplicity. In many cases it outperforms genetic algorithms. Similarly to all optimization algorithms of this type, PSO convergence relies on the use of heuristics; and convergence does not mean convergence to the optimum. In fact, the basic PSO does not guarantee convergence even to a local minimum. The basic PSO is also inefficient in dynamic optimization problems, i. e., problems in which the optimum location changes. However, variations of the basic PSO have been proven to have improved performance in dynamic problems and to be capable of convergence to local minima.

Many improvements of the PSO basic algorithm have been developed and applied successfully to a wide range of optimization problems: continuous and discrete, constrained and unconstrained, single and multi-objective, static and dynamic. Specific applications cover just about all areas of applied optimization. The main classes of applications have been in areas such as:

(1) Neural networks (training, supervised and unsupervised learning, architecture selection, etc.);
(2) Game learning;
(3) Clustering;
(4) Design (aircraft wings, antennas, circuits);
(5) Scheduling & Planning (maintenance, traveling salesman, power transmission, etc.);
(6) Controllers (flight path, air temperature, power stabilizers, etc.);
(7) Data mining. For more detail see, e. g., [6].

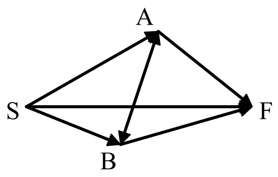
Ant Colony Optimization (ACO)

The key idea of ACO [12,13] is an abstraction and generalization of the Goss et al. [11] two-path experiment with ants. A first generalization of the two-path problem is finding the shortest path between the starting point S and the final point F when between S and F there are many possible paths.

The two-path problem can be represented by a graph with 3 vertices (S, A, F) and three arcs $(S \to A,\ A \to F,\ S \to F)$. The short path is $S \to F$ and the long path is $S \to A \to F$.

If we add another vertex B, and make the graph complete (i. e. we join with an arc each vertex to every other vertex) we obtain five possible paths: $S \to F$, $S \to A \to F$, $S \to B \to F$, $S \to A \to B \to F$, $S \to B \to A \to F$ (see Fig. 4). The idea is easily generalized to a complete graph with more vertices. The number of paths increases exponentially so checking the length of all the possible paths becomes computationally unfeasible for a large enough number of vertices.

A number N of 'ants' start at vertex S choosing randomly which vertex to go next. At every new iteration, each ant decides which arc to traverse next, with probability proportional to the amount of pheromone on the arc relative to the total amount of pheromone on the possible arcs that the ants could choose.



Swarm Intelligence, Figure 4
The five possible paths from S to F: $S \to F$; $S \to A \to F$; $S \to B \to F$; $S \to A \to B \to F$; $S \to B \to A \to F$

After an ant, k, reaches the destination F, the length of its path L_k is remembered (if an ant reaches the destination via loops, L_k is calculated after removing the loops). The ant retraces exactly the path (without loops) and deposits pheromones on the arcs in proportion to $1/L_k$.

In this way, the marking (with pheromones) of the paths by all the ants modifies the graph so that the probability of any ant taking the shortest path at the next trip from S to F increases. Eventually all the ants will follow the same path, if the algorithm converges, and the path will be the shortest if the convergence is to the global minimum. As in most optimization methods, this is never guaranteed.

The idea of the original ACO algorithm [12] adds to the foregoing sketch of the basic model three more elements. First, each arc has an a priori propensity to be traversed (regardless of pheromone content); second, each ant keeps in memory a tabu list of arcs not to be traversed, to avoid loops; and third, the pheromones evaporate at a given rate.

ACO key insight is the application of the concept of stigmergy to stochastic optimization. The ants communicate by modifying the environments (graph) and act probabilistically on the basis of the modified environment.

Many variations of the basic ACO algorithm have been proposed and implemented. The many variations take advantage of specific knowledge about the specific problem, i. e., they use heuristics, e. g., by setting the a priori propensity of traversing an arc, or by setting the evaporation rate.

ACO eventually resulted in a meta-heuristic which is a strategy for designing ACO heuristics. Various ACO-based metaheuristics have been developed. Similarly to PSO, ACO algorithms have been applied to all the basic types of optimization problems: continuous and discrete, constrained and unconstrained, single and multi-objective, static and dynamic. The first application of ACO was to the Traveling Salesman Problem, which is an NP -hard combinatorial optimization problem, and it is the most frequently attacked problem using various ACO heuristics. The main classes of other applications are to problems of

(1) Ordering (scheduling, routing);
(2) Assignment (Neural network training, image segmentation, design);
(3) Subsets finding (maximum independent set);
(4) Grouping (clustering, bin packing).

Clearly " swarm optimization" successfully uses concepts from the general notion of SI, but optimization is not in itself a necessary characteristic of SI. In fact, many tasks actually or potentially carried out by swarms are not-optimal in any sense.

Non-Linear Differential Equations Methods

One fruitful approach to modeling swarms has been to treat each individual as a discrete particle. These "individual-based" models have been employed in quite a few biological and mathematical studies. They are based on simple rules of motion for each individual, involving some combination of self-propulsion, random movement, and interaction with neighboring organisms. The models typically take the form of coupled ***non-linear*** difference or differential equations, which may be stochastic or deterministic, depending on the particular features of each model. Numerical simulations have revealed collective behavior. But a main disadvantage of such models is that, for realistic numbers of individuals, analytical results for the collective motion are difficult or impossible to obtain. It is worth mentioning that some progress has been made in obtaining analytical results for stationary groups. In [41], a discrete model was formulated, and a Lyapunov functional was used to successfully predict an equilibrium state of equally spaced organisms. However, analytical (nonstatistical) descriptions of nonequilibrium states in discrete swarm models are few.

Other investigations of swarming have been carried out in a continuum setting, in which relevant quantities are described as scalar or vector fields. This approach goes back to 1980; reviews are provided in [42]. Continuum models may be constructed a priori or by coarse-graining a particle model. In general, continuum models provide a convenient setting in which to study large populations, since one may apply machinery from the analysis of *partial differential equations*. In the context of swarms, the focus has generally been on models in which the population density satisfies a convection-diffusion equation ensuring that the population density is conserved while individuals travel with a set average velocity. Recent models of this type [43] can predict, e. g., whether a population aggregates or disperses, the regions of aggregation, and length scales of the density patterns.

Limitations of Standard-Mathematics Methods

In describing the SI investigations in the previous two sections, we have encountered the concepts of *optimization* and *non-linearity* (and earlier, in Sect. "Artificial Life Systems", *complexity* and *self-organization* arising from nonlinearity). To see more precisely the relation of these four concepts to SI we refer back to the definition of Elementary Swarm (Definition 2). Note that, by this definition, the Swarm is in principle capable of optimization, complexity and self-organization. In fact, it is clear that a Swarm is a *self-organizing* system, by definition, and that, depending on the choice of the updating function f, the pattern formed by the swarm components might, in principle, achieve high *complexity*.

As for non-linearity, it is convenient to think of f as a function of a function,

$$f = f(g(K_i)$$

where g represents the dependence from the neighbors. While g maybe linear or non-linear, the function f, representing the mode of updating, will typically be non-linear and/or probabilistic. In spite of this, the evolution of the Swarm only in special cases can be modeled by standard *non-linear dynamics* as studied via nonlinear differential equations (as we have seen in the previous section).

In regard to *optimization*, the Elementary Swarm can be easily designed to be an optimizer. In fact, if the updating is sequential, and f and v_i are chosen appropriately, one obtains a simple PSO system.

Thus, the Elementary Swarm describes a simple but powerful system capable, in principle, of self-organizing, and to produce complex structures and optimal solutions. On the other hand, even the Elementary Swarm is more general than these properties; a Swarm is not restricted by the notions of optimization or non-linear dynamics (and self-organization or complexity tied to non-linear dynamics). All these can be properties of the swarm but none is a requirement for SI. To find out what SI can do that is beyond what we have described so far, we must look at the computational capabilities of swarms. And for this we need to look at cellular-computing methods since standard-mathematics methods are ill suited to deal with computation.

Cellular-Computing Methods

Some of the first studies in SI were based on computational models, and, more precisely, on distributed algorithms applied to robotic units [44,45,46].

In these computational models, the swarm was developed as an evolution from a distributed system of processors, as follows. In *distributed computing* the algorithms are designed for a 'static' set of processing units, where, 'static' is meant literally as 'not moving'. For illustration, if a set of CPU's, computing in a distributed way, via wireless communication, started moving around, this system would look very much like a robotic swarm. In fact, referring to the intuitive notion of SI (Definition 1), all points of the SI definition would be satisfied by such a dynamic, distributed computing system provided the CPU's had, in some sense, limited capabilities.

The main point is that this distributed computing swarm differs from the robotic, and non-robotic, swarms described in the previous sections (Sects. "Standard-Mathematics Methods"–"Limitations of Standard-Mathematics Methods") in that the *intelligent task of the swarm is now seen as a 'computation'*. And this focus on computation leads us now to consider the other broad set of techniques used in SI research, i. e. techniques based not on standard-mathematics but on *cellular computing*.

Cellular computing differs qualitatively from the standard Von Neumann computing architecture. The latter is based on one complex processor that sequentially performs, at each time step, a single complex task. In contrast, in *cellular computing*, a very large number of simple processors (cells) are the units of computation. They compute (typically) in parallel with local connections between cells. The qualification 'simple' can be made precise by requiring, e. g., each cell to be a 'finite state' machine. Cellular automata are the most obvious examples of cellular computing systems but cellular computing applies to many other systems as well [47].

By definition, then, cellular computing contains several of the features of SI. The Elementary Swarm of Definition 2 can be regarded as performing a cellular computation. And in fact, cellular computing has been used extensively in A-Life studies, including systems with strong relation to SI [34]. Cellular computing systems offer SI something that the SI systems described in the 'standard-mathematics' Sects. "Standard-Mathematics Methods"–"Limitations of Standard-Mathematics Methods") lack, i. e., a clear characterization of *intelligent task*.

Intelligence as Universal Computation

Intelligence is an ambiguous concept, escaping a unique definition [48]. By identifying 'intelligence' with 'computation', the concept is restricted, but, at the same time, it can be made precise. In fact, in SI we define intelligence unambiguously as the '*ability to carry out universal computation*'.

Universal computation (or universality) is the property of a computer system (or language) which, with appropriate programming, can be made to perform exactly the same set of tasks as any other computer system (or language).

Universal computation (i. e., the ability to emulate a universal computer), is essentially the limit of any model of computation (Church–Turing thesis) [49]. It was first proven by Turing in 1936 that no system can ever carry out explicit computations more sophisticated than those carried out by a Turing machine. Subsequently, universality has been found to be a widespread property of many cellular computing systems [50].

One of the first cellular computing systems shown to be capable of universal computation, is Conway's game of life [51]. This CA is also the prototypical example of A-life systems. And it is also an example of the strong connection between universal cellular computing and bio-inspired systems.

More recently, a large number of simple cellular computing systems have been found to be capable of universal computation [50]. Many of these systems are CA, or related systems, using very simple rules of evolution with local interactions. And so they are useful starting points for modeling SI.

In particular, cellular computing is the most appropriate to endow the swarm with the property of *unpredictability*. The latter property was an original motivation for SI [1,2] and it is crucial in the task of escaping detection by a predator; it is also of importance in engineering swarms for strategic defense applications.

Unpredictability is almost a built-in property of cellular computing systems because, if one observes the rules of evolution in their raw form, it is usually almost impossible to tell much about the overall behavior they will produce.

Relations to Standard-Mathematics Methods

SMm cannot provide the swarm with the element of universal computation, which we have taken as the working definition of 'intelligence'. The only way would be to make each unit a von Neumann (i. e. standard) computing system. In a sense, this violates the notion of Swarm, since in a Swarm, by definition each unit must be 'simple'. (This point will be further clarified in Sect. "Swarms of Intelligent Units"). On the contrary, the main advantage of cellular computing systems over standard mathematics systems is the possibility of universal computation by simple units.

For this reason CCm are the natural paradigm for the understanding and designing SI systems, in spite of the fact that the approach to SI based on cellular computing has so far produced fewer so-called SI applications than the approach based on SMm. Indeed, SMm have basic limitations for modeling SI. This is because the use of SMm tends to restrict the range of tasks performable by the Swarm. And this happens because SMm typically solve problems by specifying *constraints*, i. e., conditions to be satisfied by the solution, e. g., by specifying equations. But most computational problems cannot be solved in this way.

The optimization methods described in the previous sections (PSO, ACO, etc.), illustrate the point. In these iterative methods, the key issue is what kind of changes should be made at each iteration step. Starting from a random pattern, at each step a change is made to get the pattern closer to satisfying the constraint(s). Since direct methods (e. g., gradient descent) rarely work as the pattern gets stuck into local minima, randomness in updating is added. In this way, larger portions of the solution space are sampled. The larger the changes made, the faster one can potentially approach a global minimum, but the greater the chance of overshooting. The result is that no iteration technique of this type can guarantee a solution to general combinatorial optimization problems. As we have seen, the swarm optimization methods (ACO, PSO) rely on heuristics to adjust the search and obtain often (non-optimal but) satisfactory solutions. But, in general, for the great majority of combinatorial optimization problems (e. g., the Traveling Salesman Problem [52]), no polynomial upper bound on the time complexity has been found so far. And this happens in many problems whose solution is sought by using randomness to satisfy the imposed constraints. As an example, a set of identical balls cannot be shaken into an ordered, closed-packed configuration. With extremely high probability, they lock into some configuration or another, not the optimal (close-packing) one.

This fact has *important implications* for SI. What it says is that no matter how much randomness is added to the system, it may never evolve to reach the solution specified by the constraints. Although, ultimately, constraints can be set up as a way of specifying algorithms, and hence computing, it is far simpler to specify algorithms via rules of evolution, as it is done in cellular computing.

The conclusion is that methods based on constraints and other SMm are not ideally suited for systems evolving with great complexity, and in particular they are not suitable for universal computation. Thus, if SI is to be a framework for (biological or engineered) swarms to carry out 'intelligent tasks' with the greatest generality, a methodology that allows for the swarm to carry out universal computation is necessary. To this aim, CCm are the most suitable.

Unfortunately, although CCm have many advantages over SMm for modeling SI, they address only three of the four key elements of the notion (Definition 1) of SI ('intelligence', 'collectively', 'randomly', 'unknowingly'), leaving out the element of 'randomness'. Generally, CCm operate deterministically and do not include 'randomness', as, e. g., 'swarm optimization' systems do. But this does not have necessarily to be the case. The issue is addressed in the next section.

Randomness in Swarm Intelligence

Randomness is a key element in the notion of SI (cfr. Definition 1 and Definition 2). Examples from biology justify this requirement. Randomness is not easily quantified precisely but, whatever the form and measure chosen, the point is that for swarms some form of randomness is necessary – otherwise they would fail to be models for analyzing a large class of biological systems. But what kind of randomness is essential to model these biological systems?

Randomness in the number and type of agents is not important – the agents could be strictly identical and remain in the same number. Randomness in the initial conditions is not essential either. Many swarms evolve from regular initial conditions into highly complex and random patterns. Randomness of external input from the environment is not always present, and it is certainly not a requirement for biological swarm behavior.

What about the randomness artificially added to the units, as in swarm optimization?

The randomness added to the units in PSO or ACO algorithms is modeled as originating from the random behavior of each unit. This is a plausible assumption in relation to biological systems. But the swarms in PSO and ACO are updated in an orderly (non-random) way, typically sequentially (there are also some parallel implementations [13]) whereas, in biological systems, the units update in a disordered, random fashion.

And it is this type of randomness that is both necessary in any biologically relevant model of swarms and sufficient to provide many (but not all) of the advantages of randomness in solving swarm engineering problems.

The conclusion is that the only randomness that is truly essential for SI is randomness in the times of operation of the units. Each unit has its own clock, not synchronized with other units' clocks. Other types of randomness in the behavior of the units or the environment may be required to solve specific problems, but randomness in times or operation is necessary for any biologically realistic model.

Interestingly though, many applications so far considered in the area of SI do not yet include this randomness in the models. We have already mentioned that typical optimizing swarms update sequentially; and CA systems operate largely in parallel, i. e. synchronously. Synchronous or sequential operations are by far the most common updating modes in either SMm or CCm.

The Implicit Assumption of Asynchrony Irrelevance

As noted, it is a basic fact that biological agents, apart from exceptional cases, do not operate synchronously (or sequentially) in groups. It is also a fact that people in social groups do not operate synchronously, or sequentially. If SI is supposed to model biological and social swarms, SI must be based on models that do not operate synchronously or sequentially [53].

And if biological swarms are capable of solving problems (including optimization) without synchrony (nor sequentially), as they do, then models that imitate those swarms should operate asynchronously (not sequentially).

But, as noted, the main modeling paradigms for bio-inspired algorithms, standard-mathematics and cellular computing, are either essentially sequential or synchronous.

An example from SMm is the solution of partial differential equations: they operate synchronously on every point (clearly seen in solving them numerically and iteratively). This unrealistic use of differential equations in biological processes has been pointed out, e. g., in the problem of morphogenesis [54]. The Turing diffusion-reaction model [55], being based on differential equations, implies synchronicity and central control, hence it is physically not realistic for a scale of the order of 100 cells.

In fact, synchronicity leads to realistic models only whenever the spatio-temporal resolution is high, as, e. g., for phenomena typically studied in physics. But when the units studied are complex or few enough to have a less fine spatio-temporal resolution, as in biology or human societies, synchronicity is not realistic, as it is obvious by observation.

Thus, in using synchronous methods for biological or human societies, implicitly a strong assumption is being made, i. e., that the synchronously (or sequentially) and non-synchronously (and not sequentially) obtained solutions would coincide.

But this assumption has no validity. In fact, it has been shown, for example, that CA, when running in synchronous and non-synchronous ways, normally produce totally different results [56]. This has been noted already in the nineties [57] in A-Life studies. In [57] two well-known CA were compared: Conway's "game of life" [52], and the Immune Network model. The former is a 2-dimensional CA capable of universal computation when run synchronously. But the behavior is totally different when run without synchrony: the Game of Life stops producing complex patterns and converges to a fixed point.

The Immune Network model is asynchronous and the Game of Life synchronous. The crucial factor in the different behavior of the two systems was identified as the synchronous vs. asynchronous updating. In fact, it was concluded that, in this case, asynchrony induces stability in CA. This agrees qualitatively with studies in standard-mathematics [58].

In conclusion, the assumption that asynchrony makes no difference has been found not to be valid. Hence, asynchronous systems must be studied as such, not by using synchronous models. Moreover, different types of asynchrony yield different results, as discussed in Sect. "Asynchronous Swarms".

Asynchronous Swarms

Several cellular computing studies in the nineties [37,59] led to a variety of results emphasizing the role that different types of asynchrony play in the results. Studying asynchronous systems is complicated because, among other things, deviation from synchronicity, i. e., from the mode of updating all units in parallel at each time step, may occur in several different ways. For example, sequential updating and random updating are both asynchronous but very different.

Types of Asynchrony

Unfortunately, there is no standard vocabulary for the various types of asynchrony. Thus we use the following classification to describe the possible types of asynchrony.

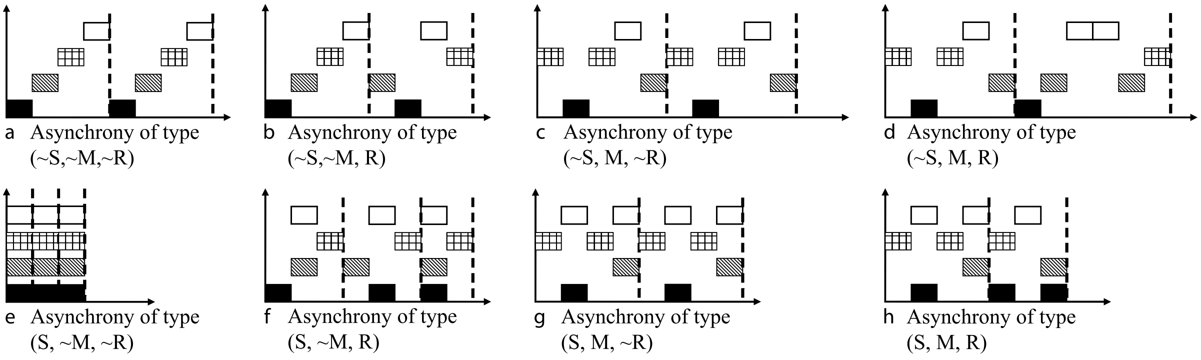
Consider an updating cycle (UC), i. e., the time interval at the end of which all units have been updated at least once. Eight types of UC can be identified by the presence or absence of any of following three properties: (*S*) *Synchronicity*: more than one unit may update at each time step; (*M*) *Multiplicity*: any unit may update more than once in each UC; (*R*) *Randomness*: the updating order varies randomly for every UC (see Fig. 5).

These eight basic types of asynchronous updating, can be further specialized. For example, if all three properties are absent ($\sim S$, $\sim M$, $\sim R$), the updating is sequential. But the sequential updating order of the units can be fixed in different ways. Studies of CA have proven that the behavior differs markedly not only for the eight types of asynchrony, but even among different sequential ordering [37,56].

In [56] the (S, M, $\sim R$) form of updating has been applied to describe processes where each unit has independent clocks but the clocks have a fixed, non random frequency. This type of asynchrony is considered a good model for forest ecosystems, fire spread, and other natural and artificial systems. The results are very different when updating of the type ($\sim S$, M, R), ($\sim S$, $\sim M$, R), or sequential ($\sim S$, $\sim M$, $\sim R$) are applied to the same system.

In conclusion, the crucial point is that [56] the exact manner of updating can have a profound effect on overall system behavior. The implication of this is that when comparing models of natural systems or artificial multi-agent systems it must be stated which updating scheme has been used, otherwise meaningful comparison between different studies may not be possible.

In particular, returning to swarm optimization, one may ask whether in tasks such as finding the shortest path, it is realistic to apply to natural systems (such as in-


a Asynchrony of type (~S,~M,~R)
b Asynchrony of type (~S,~M, R)
c Asynchrony of type (~S, M, ~R)
d Asynchrony of type (~S, M, R)
e Asynchrony of type (S, ~M, ~R)
f Asynchrony of type (S, ~M, R)
g Asynchrony of type (S, M, ~R)
h Asynchrony of type (S, M, R)

Swarm Intelligence, Figure 5
a. Asynchrony of type ($\sim S$, $\sim M$, $\sim R$); b. Asynchrony of type ($\sim S$, $\sim M$, R); c. Asynchrony of type ($\sim S$, M, $\sim R$); d. Asynchrony of type ($\sim S$, M, R); e. Asynchrony of type (S, $\sim M$, $\sim R$); f. Asynchrony of type (S, $\sim M$, R); g. Asynchrony of type (S, M, $\sim R$); h. Asynchrony of type (S, M, R); Illustration of the 8 types of updating, according to Synchronicity, Multiplicity, and Randomness. Four units are represented by rectangles with different patterns, from black (*bottom*) to white (*top*). The horizontal axis measures time steps in units equal to the base of a rectangle. The vertical dashed line indicates the end of an updating cycle (i. e., all units have updated at least once). The label below the horizontal axis specifies the type of updating (∼ means 'not'). The standard 'parallel' and 'sequential' updating are, respectively, (S,$\sim M$,$\sim R$), Fig. 5e, and ($\sim S$, $\sim M$, $\sim R$), Fig. 5a

sect societies) swarm models which are sequential (such as swarm optimization models) or synchronous (such as models based on differential equations). While these models work effectively as artificial swarms there is no proof that they apply to natural systems, which are asynchronous.

Modeling Asynchrony by Synchronous Swarms

Because of the widespread use of synchronous methods in simulations of SI, one might wonder under what conditions a synchronous but stochastic model could be equivalent to an asynchronous one.

To answer this question, let us consider the two types of stochastic models most commonly used to model randomness in synchronously updated systems. The randomness may be included in (1) the possible outcomes of the updating function or (2) in the choice of the function applied to the updating.

Referring to the definition of Elementary Swarm (Definition 2) the two cases correspond to generalizing the updating function as follows:

$$\text{Case (1)} \quad \forall i \in N\colon v_i(t+1) = f(\mathbf{v}_{k \in K(i)}(t)\,\zeta)$$

where ζ is a random variable.

$$\text{Case (2)} \quad \forall i \in N\colon v_i(t+1) = f_{(t)}(\mathbf{v}_{k \in K(i)}(t))\,;$$

where $P[f_{(t)} = f_\gamma]$ is the probability mass function of choosing $f_{(t)} = f_\gamma$ out of a set of N_f possible functions $\{f_\gamma; \gamma = 1, \ldots, N_f\}$.

Case (1) is typical of probabilistic CA, and it is also the method used in PSO. In these systems the state vector, at each time step, evolves according to a fixed rule which produce a new state vector from the previous one. The rule is based on the state of the neighbors of each unit and does not change from step to step but the outcome of the rule is probabilistic.

Case (2) is what is done for example in probabilistic Iterated Function Systems [23]. In probabilistic Iterated Function Systems a vector evolves via a set of maps (a map is a function whose domain and range coincide); at each time step a map is chosen, probabilistically, from a set of possible maps.

In either cases (1) or (2), the updating scheme fails to model the actual time evolution of natural systems not so much because the updating are applied synchronously but because the randomness is applied *collectively*, i. e. to all the units in the same way. On the other hand a synchronous algorithm realistically simulating independent random updating can be run as case (2) applied *individually* to each unit, as follows:

$$\text{Case (3)} \quad \forall i \in N\colon v_i(t+1) = f_{(t)i}(\mathbf{v}_{k \in K(i)}(t))$$

where $P[f_{(t)1} = f_{\gamma 1},\ f_{(t)2} = f_{\gamma 2}, \ldots, f_{(t)N} = f_{\gamma N}]$ with $f_{(t)i} \in \{f_\gamma; \gamma = 1, \ldots, N_f\}$ is the joint probability mass function of each unit i updating, at time t, according to the function $f_{(t)i}$.

In the simplest embodiment of case (3), the set of possible updating functions consists only of the identity and of another function f, with probabilities p and $(1-p)$ respectively (i. e., a Bernoulli process). In such a case, every unit, at each time step, either does not update, with probability p, or updates according to the function f, with probability $(1-p)$. Running this algorithm synchronously is equivalent to asynchronous independent updating of the units in a random way – a realistic description of a random swarm. So, under these independently stochastic conditions, running a simulation synchronously, represents correctly the physical asynchronous updating of the swarm units. On the other hand, this does not change the fact that different results are obtained when using this random updating (whether simulated with stochastic synchrony or not) instead of synchronous or sequential updating.

Local Synchrony and Self-Synchronization

Another approach to dealing with the problem of random updating by the swarm units, is to explore the possibility of self-synchronization. If the swarm can self-synchronize, then all the results for synchronous swarms could be applied.

To look into this, let us return to the classification of the types of asynchrony, i. e., the SMR classification above. If the SMR properties are applied to blocks of units rather than individual units, the resulting updating orders are referred to as '*locally synchronous*'.

CA with cells organized into blocks have been investigated [60]. These CA relax the normal requirement of all cells having the same update rule. Cells within a block are updated synchronously, but blocks are updated asynchronously. They experimented with different SMR types of asynchrony and concluded that synchronous and asynchronous CA can be evolved with equivalent computational properties, but CA of the asynchronous type may require a larger number of cells [60]. Another study [61] has shown cases in which local synchronization can lead to the same outcome as with global synchronization. But how can local synchronization be achieved?

A number of schemes have appeared in which the order of updating depends on local interactions and leads to local synchronization. In effect, what local interactions (or constraints) can do is to force a unit to wait to update until others are ready, and so this creates a local synchronization. An asynchronous CA model that can behave as a synchronous CA has been demonstrated [62]; it functions by the addition of extra constraints on the order of updating, effectively providing a type of local synchronization. Whether these methods of self-synchronization may in some cases result in realistic models of natural systems of SI remains an open question.

The Natural Asynchrony of Swarms

We have seen that the implicit assumption of asynchrony-synchrony equivalence must be rejected and that different types of asynchrony give different results. But what type of asynchrony is most relevant to SI? There is no easy answer. For example, ants work and rest; active and resting periods have an aperiodic pattern for individual ants, but for the whole colony there are synchronized periodic patterns of active and resting periods.

In spite of the difficulty of finding a clear cut answer to the question of the natural mode of SI updating, from observations of biological systems, and from local synchronization models, it may be plausible to assume that the essential form of asynchrony in SI is the randomness in the working of the individual clocks, as argued in Sect. "Randomness in Swarm Intelligence"; hence the SI asynchrony must be characterized by the presence of all three asynchrony properties, i. e., (SMR).

In conclusion, at this stage of our discourse, the Swarm remains defined as in Definition 2, qualified by SMR asynchronous updating, which hereinafter we call *'natural'* asynchrony. Note that stochastic synchronous simulations of this model can also be carried out as, e. g., in case (3) above.

The Realization of Asynchronous Swarms

So far we have established the importance and type of asynchronous models in SI, but what SI investigations using asynchronous swarms are there?

As noted in Sect. "Asynchronous Swarms", research in asynchronous models is still very limited, relative to synchronous models, and this in spite of the fact that the very first models of SI were all asynchronous, using SMm based on finite differences [46]. Explicit updating schemes in finite difference methods can also be regarded as parallel CA, thus belonging to both SMm and CCm. Investigations of asynchronicity in finite difference methods are not common [58]. Examples include a non-linear updating rule was based on a linear relation between two neighboring units [45]. A gradient type of swarm updating, was also proposed in modeling morphogenesis [54].

For swarms updating with 'natural' asynchrony, i. e., according to (SMR), a recent study [58] gives a proof of convergence to the same state as by using synchronous or sequential iterations. It was also shown that, under certain conditions, the (SMR) asynchronous updating leads to convergence while synchronous updating does not. This is another example of the advantages of randomness in allowing the swarm to reach a fixed state.

At the end of Sect. "Cellular-Computing Methods" we concluded that CCm have, for SI modeling, many advantages over SMm. The most crucial advantage is the possibility of universal computation which we took as the definition of intelligence for SI. We also noted, however, that studies based on CCm which include randomness are scarce. We described a few in Sect. "Randomness in Swarm Intelligence", especially in discussing the qualitative differences with synchronous CA and in relation to mechanisms of local and self-synchronization.

Generally these studies model relatively trivial phenomena but cannot model nontrivial phenomena such as universal computation. In fact, there are very few studies of universal computation in asynchronous CA. Significant advances have been made only recently. The first attempts were made by simulating a synchronous CA on an asynchronous CA [63] after which a synchronous model, as a Turing Machine, was simulated on the synchronous CA. However, this asynchronous CA is, in practical realization, synchronous.

Improved asynchronous CA do not rely on global synchronization but conduct asynchronous computation directly by simulating Delay-Insensitive circuits, i. e., circuits in which delays of signals do not affect the correctness of the circuit operation [64]. This method essentially uses local synchronization with undetermined exact timing between transitions. In this way an asynchronous CA, with a hexagonal cell structure, capable of universal computing has been realized [65]. Although relying on local synchronization, this type of asynchronous CA can mimic natural phenomena as, e. g., phenomena that rely on chemical reactions which occur only when the right molecules are available in the right positions at the right times.

More recently a computation-universal and construction-universal asynchronous CA has been designed [66] and used to implement self-reproducing [67,68] machines. Besides computational universality, construction universality is important in SI because it allows the swarm to be hardware reconfigurable, an important characteristic of many biological systems.

We note that the recent interest in asynchronous CA stems not directly from SI but from nanotechnology. In fact nanocomputer architectures with asynchronous updating may reduce heat dissipation, an important limiting factor in scaling down the size of computing chips [64]. In this respect, it is likely that SI concepts will play a major role in nanoscale systems.

In conclusion, the recent realization [66] of universal asynchronous CA is a major step toward the realization of true SI.

Characteristics of Swarm Intelligence

The demonstration of universal computation in asynchronous CA, amounts to a validation of the concept of SI. In fact we can now combine universal computation with the Elementary Swarm definition (Definition 2) to quantify the intuitive definition of SI (Definition 1), as

Definition 3 SI is the study of *universal cellular-computing systems updating with natural asynchrony*.

Here 'natural' means SMR-asynchrony (see Sect. "Asynchronous Swarms") or updating randomly in parallel, as in case (3) of Sect. "Asynchronous Swarms". We call it 'natural' since it models the natural mode of updating of typical biological swarms and human societies.

A few remarks about Definition 3. The four elements of the intuitive notions (Definition 1) are made precise by Definition 3: 'collectively' and 'unknowingly' are inherent in the structure of cellular computing; 'intelligence' is in universal computation; and 'randomness' is in the natural asynchronous operation.

Definition 3 deals with CCm and may appear to exclude SMm in SI; but this is not the case. In fact, many SMm, as used in SI, can be regarded as special cases of cellular computing methods, as, e. g., are swarm optimization and iterative methods in finite differences.

Also Definition 3 does imply that every SI system must be capable of universal computation; what Definition 3 does is to establish a focus of attention for the SI area of studies and at the same time give a precise and realistic meaning to the kind of 'intelligence' aimed at in SI, rather than the often vague and exaggerated meanings given in the popular literature.

Definition 3 also indicates how SI becomes of relevance beyond biological systems and robotics. In fact SI will likely be an important concept in the future of computation. At very small scales, time delays between computational components cannot guarantee synchrony; the various components must have independent clocks, thus beginning to resemble the operation of a biological swarm. So, swarms are likely to be studied extensively in connections with nanocomputing.

The above arguments bring up the question as to whether SI is nothing more than asynchronous cellular computing. The answer is that designs of asynchronous CA, as investigated in computer engineering, are generally not models of SI. A first, basic reason is that, with one recent exception [66], asynchronous CA, do not update 'naturally'. A second, more fundamental reason, is that in most SI studies (both in natural and technological systems) **the units are dynamic**. When the units of an asynchronous CA are made mobile, a different, and more complex set of problems need to be solved due to the changing neighborhoods of each cell. These issues of dynamic reconfigurations of cells have not been addressed in asynchronous cellular computing designs and are likely to remain outside the scope of research aimed at improved computer architectures. The computational problems arising from dynamically reconfiguring cells are central in SI. We address this issue next.

Dynamics in Swarm Intelligence

In the definitions of SI and Swarm given so far (Def. 1., Def. 2., Def. 3.), there has been no mention of the dynamics of the units. But a general characteristic of the units of a swarm is that almost invariably they are mobile. In fact we have already discussed the dynamic nature of swarms in Sects. "Biological Systems" and "Robotic Systems" in relation to biological and robotic swarms. The reason we have so far omitted this dynamic character of the units from the progressively more precise definitions of SI is simply for clarity of exposition: if the dynamics is introduced after all the other elements of SI have been defined and quantified it is easier to single out its real importance.

From Definition 3, with appropriate specializations, all aspects of SI considered so far can be included in a common core of studies. The most general notion of SI is in fact that of *universal computation carried out with 'natural asynchrony' by a cellular computing system*. But we should add, *whose cells are, in general, mobile units*.

The latter qualification would be unnecessary if the description of the dynamic state could be included among the state variables of the cell. But this is not possible in computing cells since the computation depends on neighbors that change their locations. The fact that *dynamic cellular computing systems* (also called *cellular robotic systems*) are not equivalent to (and very hard to simulate by) cellular computers has been emphasized since the very beginning of SI studies [44]. Thus, we conclude by stating a definition of SI which, while remaining grounded in the intuitive ideas (Definition 1), includes all the concepts discussed quantitatively in this article.

Definition 4 SI is the capability of universal computation carried out with 'natural' asynchrony by a dynamic cellular computing system.

As we have seen, studies in the area of SI so far have been concerned with models of collective behavior which, to some limited degree, approach SI as defined above. Even though, to date, no system with SI (Def. 4.) has been built (or designed), significant progress has been made and, from what we have seen, it is reasonable to expect that it can be done.

In fact, although there is not yet proof of universal computation carried out with 'natural' asynchrony by a *cellular robotic system* (i. e., a dynamic cellular computing system), the recent proof [66] for 'static' cellular computers indicates that this may be possible in the near future. Other future perspectives for SI are discussed in Sect. "Future Directions".

Unpredictability in Swarm Intelligence

We are now in a position to consider an aspect of SI which has been inherent to the concept of SI from its inception, i. e., the unpredictability [1,2] of the Swarm.

The unpredictability of the Swarm agrees with the common intuition that it is usually difficult to predict what a program will do by reading its code, and the more so the lower the level of the language used by the program. More precisely, a Swarm, like other universal computers, may be impossible to predict in the sense that even if one knows the rules of evolution and an initial state, it can still take an irreducible [50,69] amount of computation to actually predict future states. Furthermore, the unpredictability of the Swarm is of a more general character than that of any universal computer because of the randomness inherent in its evolution and because of its dynamics.

The unpredictability of a Swarm by a Von Neumann universal computer has been argued in [1,10] on the basis of its dynamics. The unpredictability of a Swarm by a cellular automaton has also been discussed [1,10]. Although unpredictability is difficult to quantify, it is generally engineered by adding randomness to the system as in camouflage and cryptography. Also, in animals, randomness and dynamics are used by a herd to avoid predators by becoming unpredictable. And so in team sports, such as soccer, unpredictability by the opponent is usually achieved by a combination of randomness and dynamics.

Therefore, intuitively we may conjecture that among systems capable of universal computation, a Swarm, because it operates with randomness and dynamically, would be *the least predictable*.

Swarms of Intelligent Units

We can now consider collections of intelligent units, i. e., such that each unit is capable of universal computation. These seem excluded from the definitions of SI given so far – a key characteristic of SI is that intelligence is an emergent property, happening only above a certain critical number of units, and not a property of any of the individual units. On the other hand, some of these groups are often included in broad considerations of SI [4] as applied to human societies. Under what conditions can these groups be regarded as swarms?

A simple answer to this question runs as follows. Consider the special case of Definition 4, when the cells are not finite state machines but universal computing units. As long as the task at hand cannot be accomplished by a single unit, but only by more than a critical number of units, the system operates as a swarm. That this can be the case is supported intuitively and from computational considerations, as follows.

Intuitively, we may refer back to the example of the free market model mentioned as illustration of SI at the beginning of the article. Each individual contributes only as a trader; the 'computation' of the market price is done by the swarm collectively and could not be done by any individual agent. The point is that each unit, albeit 'intelligent', uses only a fraction of its capability, i. e., the trading ability, thus operating, effectively, in a restricted, 'non-intelligent' capacity.

Computationally, we have noted in discussing unpredictability that, in spite of theoretical computational equivalence among universal computers, the capability for one universal computer to predict another is limited. And, in fact, at the end of Sect. "Swarms of Intelligent Units", we put forth the conjecture that a Swarm is the least predictable universal computer.

But even if this conjecture were not true, it still makes sense to think of a case when a Swarm of universal computers is unpredictable by any of the units comprising it, as it has been argued for the case of units capable of universal computation with Von Neumann architecture [1,10]. In this sense we may think of a swarm of universal computers as being more capable than any one of its units, in spite of the fact that the Swarm and any of its units are computationally equivalent. For these reasons, it makes sense to apply, under appropriate conditions, the notion of SI also to human societies.

We may give then a more complete definition of Swarm by adding the characteristic of unpredictability of the Swarm by its units.

Definition 5 SI is the capability of universal computation carried out with 'natural' asynchrony by a dynamic cellular computing system, none of whose cells can predict the computation done by the Swarm.

The latter specification is obviously redundant for common cellular computing systems but it is useful to exclude from SI trivial cases of human social activities; for example, activities of human groups whose association cannot be proven to solve a problem that could not have been solved by any individual alone.

Many popular interpretations of SI have appropriated the label SI to refer to almost any trivial human group activity, such as brainstorming. In such activities often there is no way of proving that the final output of the group could not have been predicted by one of the members of the group. Ultimately, however, these considerations require an understanding of the relation between human thinking and computation, and thus fall beyond the scope of this article.

Similarly, beyond the scope of this article and SI, fall studies of multiagent systems in Artificial Intelligence [70]. In the Artificial Intelligence area, the emphasis on multi-agent systems is in finding decision algorithms, i. e. 'agents', for open environments in which these agents must operate robustly and rapidly, i. e. 'intelligently'. Generally, however, the problem of collective decision making, organization theory, distributed reasoning and distributed Artificial Intelligence is typically beyond the scope of SI.

SI deals with human groups only when they operate at a low level of intelligence. Definition 5 includes human groups of individuals that operate under restrictions which the Swarm can overcome. An example is that of groups of individuals each with a limited computation device connected wirelessly to neighbors as in a cellular computing system. With appropriate algorithms such a Swarm could compute universally while the individual device cannot. This concept is applicable in, e. g., defense operations or emergency mass evacuations strategies.

In conclusion, Definition 5 embodies the intuitive definition of SI (Definition 1) and indicates why SI methods can be used to solve problems not solvable by traditional methods. Besides having all the properties of universal CA, the Swarm operates, as natural systems do, by independent clocks with no centralization and can be designed to be dynamic and unpredictable by any system including any its own units.

Future Directions

The field of SI is still less than 20 years old and it is in a formative phase, with SI researchers engaged in a broad range of disciplines. During these formative years the 'conversation' about SI has followed several strands around some common themes with various emphases ranging from speculative inquiries to practical interests, as we have seen in this article.

The meaning of the term SI has tended to broaden, covering now many areas, apparently only weakly related by the some intuitive notions. We have seen, however, how all these SI ideas have a 'center of attraction' in a basic concept of SI that can be made precise and quantified (Definition 4 or 5), and thus used to provide unity, continuity, and boundaries, thus preventing the area from broadening to the point of being unable to sustain an effective research community.

With this perspective, SI can also be seen as having an ultimate theoretical goal for the practical realization of engineered Swarms, whether robotic, biological or simply computational.

Practically, the goal of SI will remain two-fold: to provide models to explain biological societies and to engineer algorithms and devices with capability beyond those of traditional technologies. It will continue to include Swarm Robotics and bio-inspired algorithms such as swarm optimization methods.

But although commonly regarded as a typical example of bio-inspired technology, SI applications are likely to go beyond bio-inspired systems. We can see this if we consider how nature-inspired technologies have evolved.

Science discovers laws of nature, and technology makes inventions using those laws, often together with design ideas also derived from nature. Thus, for example, laws of physics and designs inspired by crystal structures are now applied to nanotechnologies; similarly, laws of biology and designs inspired by genetic configurations are applied to make artificial organisms in biotechnology. More recently a new kind of science [50] is discovering the laws and designs of computing machines as though they were natural systems, and these discoveries are likely to be used to invent new software algorithms and hardware implementations of those algorithms. SI, besides being bio-inspired, can be said to be inspired also by this new science, which, in some respects, can be more general than biology.

In this perspective, SI will evolve into the study of what amounts to be very powerful computing systems. Designing the simplest of these, i. e., the simplest universal dynamic cellular-computing system updating with natural asynchrony, is an example of a future theoretical and practical challenge for SI. More immediate, future applications can be extrapolated from the examples given throughout the article.

Bibliography

Primary Literature

1. Beni G, Wang J (1989) Swarm Intelligence in Cellular Robotic Systems, Proceed. NATO Advanced Workshop on Robots and Biological Systems, Tuscany, Italy, June 26–30
2. Beni G, Wang J (1989) Swarm Intelligence. Proc. 7th Ann Meeting of the Robotics Society of Japan, pp 425–428 (in Japanese)
3. Bonabeau E, Dorigo M, Theraulaz G (1999) Swarm Intelligence: From Natural to Artificial Systems. Oxford Univ Press, New York
4. Kennedy J, Eberhart RC, Shi Y (2001) Swarm Intelligence. Morgan Kauffman, San Mateo
5. Abraham A, Grosan C, Ramos V (2006) Swarm Intelligence in Data Mining (Studies in Computational Intelligence). Springer, vol 34
6. Engelbrecht AP (2006) Fundamentals of Computational Swarm Intelligence. Wiley, New York
7. Dorigo M, Gambardella LM, Birattari M, Martinoli A (eds) (2006) Ant Colony Optimization and Swarm Intelligence: 5th International Workshop, ANTS 2006, Brussels, Belgium, September 4–7, (2006), Proceedings (Lecture Notes in Computer Science). Springer, Berlin
8. IEEE Swarm Intelligence Symposium, Honolulu, Hawaii, 1–5 April 2007. http://www.computelligence.org/sis/2007/?q=node/2. Accessed 6 Mar 2008
9. Dorigo M (ed) Swarm Intelligence, ISSN: 1935-3812. Springer US (to begin publication Summer 07)
10. Beni G (2004) From Swarm Intelligence to Swarm Robotics. In: Sahin E, Spear WM (eds.) Swarm Robotics March, Revised Selected Papers, SAB 2004 International Workshop, Santa Monica, CA, 17 July 2004. Lecture Notes in Computer Science vol 3342. Springer, Berlin, pp 1–9
11. Goss S, Aron S, Deneubourg JL, Pasteel JM (1989) Self-organized shortcuts in the Argentine ant. Naturwissenschaften 76:579–581
12. Dorigo M (1992) Optimization, Learning and Natural Algorithms. Ph D Thesis, Dipartimento di Elettronica, Politecnico di Milano, Milan (in Italian)
13. Dorigo M, Stutzle T (2004) Ant Colony Optimization. MIT Press, Cambridge
14. Olariu S, Zomaya AY (2005) Handbook of Bioinspired Algorithms and Applications. Chapman Hall/Crc Computer Information Science, Boca Raton
15. Passino K (2004) Biomimicry for Optimization, Control, and Automation. Springer, London
16. Reynolds C (1987) Flocks, Herds, and Schools: a distributed behavioral model. Computer Graphics 21(4):25–34
17. Moere AV (2004) Information Flocking: Time-Varying Data Visualization using Boid Behaviors. Proc Eighth Int Conf Inform Visual, pp 409–414
18. Eftimie R, de Vries G, Lewis MA (2007) Complex spatial group patterns result from different animal communication mechanisms. Proc Nat Acad Sci 104(17)
19. Parker LE, Schneider FE, Schultz AC (2005) Multi-Robot Systems. From Swarms to Intelligent Automata, vol 3: Proceedings from the 2005 International Workshop on Multi-Robot Systems. Springer, Dordrecht
20. Sahin E, Spears WM (2005) Swarm Robotics. SAB 2004 International Workshop, Santa Monica, CA, July 17, 2004, Revised Selected Papers (Lecture Notes in Computer Science) Springer, Berlin
21. Winfield A, Redi J http://www.robocomm.org/2007/index.shtml. Accessed 6 Mar 2008
22. Sahin EL, Spears WM, Winfield AFT (Eds.) (2007) Swarm Robotics. Second SAB 2006 International Workshop, Rome, Italy, September 30-October 1, 2006 Revised Selected Papers. Series: Lecture Notes in Computer Science vol 4433. Springer, Berlin
23. Baldassarre G, Trianni V, Bonani M, Mondada F, Dorigo M, Nolfi S (2007) Self-organized coordinated motion in groups of physically connected robots. IEEE Trans Syst Man Cybern Part B: Cybernetics 37(1):224–239
24. Gazi V, Passino KM (2004) Stability Analysis of Social Foraging Swarms. IEEE Trans Syst Man Cybern B 34:539–557
25. Balch T, Dellaert F, Feldman A, Guillory A, Isbell CL Jr, Khan Z, Pratt SC, Stein AN, Wilde H (2006) How AI and Multi-Robot Systems research will accelerate our understanding of social animal behavior. Proc IEEE 94(7):1445–1463
26. Brooks R (1986) A robust layered control system for a mobile robot. IEEE J Robot Autom RA-2(1):14
27. Johnson N, Galata A, Hogg DB (1998) The acquisition and use of interaction behavior models. In: Proc IEEE Computer Society Conf Computer Vision and Pattern Recognition, pp 866–871, IEEE Computer Society Press, Santa Barbara
28. Curtis SA, Mica J, Nuth J, Marr G, Rilee ML, Bhat M (2000) Autonomous Nano-Technology Swarm. Proceedings of the 51st International Aeronautical Congress, IAF-00-Q5.08
29. ANTS website: http://ants.gsfc.nasa.gov, Goddard Ants Team. Accessed 6 Apr 2008
30. Brooks RA (1991) Intelligence without representation. Artif Intell 47:139–159
31. Dorigo M, Tuci E, Groß R, Trianni V, Labella TH, Nouyan S, Ampatzis C, Deneubourg J-L, Baldassarre G, Nolfi S, Mondada F, Floreano D, Gambardella LM (2004) The SWARM-BOTS project. In: Sahin E, Spears WM (eds) Proc. of the 1st Int. Workshop on Swarm Robotics. LNCS vol 3342. Springer, Berlin, pp 26–40
32. Mondada F, Pettinaro GC, Guignard A, Kwee IV, Floreano D, Deneubourg J-L, Nolfi S, Gambardella LM, Dorigo M (2004) SWARM-BOT: A new distributed robotic concept. Auton Robots 17(2–3):193–221
33. Mondada F, Gambardella LM, Floreano D, Nolfi S, Deneubourg J-L, Dorigo M (2005) The cooperation of swarm-bots: Physical interactions in collective robotics. IEEE Robot Autom Mag 12(2):21–28
34. Rocha LM, Yaeger LS, Bedau MA, Floreano D, Goldstone RL, Vespignani A (2006) Artificial Life X Proceedings of the Tenth International Conference on the Simulation and Synthesis of Living Systems. MIT Press, Cambridge
35. Traub J (ed) Journal of Complexity. Elsevier http://www.elsevier.com/wps/find/journaldescription.cws_home/622865/description#description
36. Nicolis G, Prigogine I (1977) Self-Organization in Non-Equilibrium Systems. Wiley, New York

37. Sipper M (1999) The Emergence of Cellular Computing. IEEE Comput 32(7):18–26
38. Kennedy J, Eberhart RC (1995) Particle swarm optimization. In: Proc. IEEE int'l conf. on neural networks vol IV, pp 1942–1948. IEEE service center, Piscataway
39. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by Simulated Annealing. Science 220(4598):671–680
40. Snyman JA (2005) Practical Mathematical Optimization: An Introduction to Basic Optimization Theory and Classical and New Gradient-Based Algorithms. Springer Science-Business Media, New York
41. Mogilner A, Edelstein-Keshet L, Bent L, Spiros A (2003) Mutual interactions, potentials, and individual distance in a social aggregation. J Math Biol 47:353–389
42. Murray JD (2002) Mathematical Biology I: An Introduction, 3rd edn., Interdiscip Appl Math 17. Springer, New York
43. Topaz CM, Bertozzi A (2004) Swarming Patterns in two-dimensional kinematic model for biological groups. SIAM J Appl Math 65(1):152–174
44. Beni G (1988) The concept of cellular robot. Proc 3rd IEEE Symposium on Intelligent Control, pp 57–61. Arlington
45. Beni G, Hackwood S (1992) Stationary waves in cyclic swarms. In: Proc IEEE International Symposium on Intelligent Control. August 10–13, 1992. Institute of Electrical and Electronics Engineers (IEEE), Glasgow
46. Beni G (1992) Distributed Robotic Systems and Swarm Intelligence. J Robot Soc Japan (in Japanese) 10:31–37
47. Sipper M (1997) Evolution of Parallel Cellular Machines: The Cellular Programming Approach. Lecture Notes in Computer Science. Springer
48. Gottfredson LS (1997) Mainstream Science on Intelligence: An Editorial with 52 Signatories, History, and Bibliography. Intelligence 24(1):13–23
49. Cooper SB (2003) Computability Theory. Chapman Hall/CRC, Boca Raton
50. Wolfram S (2002) A New Kind of Science. Wolfram Media, Champaign
51. Gardner M (1970) The fantastic combinations of John Conway's new solitaire game 'life'. Sci Am 223:120–123
52. Johnson DS, McGeoch LA (2003) The traveling salesman problem: A case study in local optimization. In: Aarts EHL, Lenstra JK (eds) Local Search in Combinatorial Optimization, pp 215–310. Wiley, Chichester
53. Huberman BA, Glance NS (1993) Evolutionary Games and computer simulations. Proc Nati Acad Sci USA 90:7716–7718
54. Liang P, Beni G (1995) Robotic Morphogenesis. Proc Int Conf Robot Automat 2:2175–2180
55. Turing AM (1952) The Chemical Basis for Morphogenesis. Phil Trans Roy Soc London B 237:37–72
56. Cornforth D, Green D, Newth D (2005) Ordered Asynchronous Processes in Multi-Agent Systems. Physica D 204(1–2):70–82
57. Bersini H, Detour V (1994) Asynchrony induces stability in CA based models. In: Brooks RA, Maes P (eds) Artificial Life IV. MIT Press, Cambridge, pp 382–387
58. Beni G (2004) Order by Disordered Action in Swarms. In: Sahin E, Spear WM (eds.) (2005) Swarm Robotics, SAB 2004 International Workshop. LNCS, vol 3342. Springer, Berlin, pp 153–171
59. Schonfisch B, deRoos A (1999) Synchronous and Asynchronous Updating in Cellular Automata. Biosyst 51:123–143
60. Sipper M, Tomassini M, Capcarrere MS (1997) Evolving asynchronous and scalable non-uniform cellular automata. In: Proceedings of International Conference on Artificial Neural Networks and Genetic Algorithms (ICANNGA97). Springer, New York
61. Clapham N (2002) Emergent synchrony: simple asynchronous update rules can produce synchronous behavior. In: Sarker R, McKay RI, Gen M, Namatame A (eds) Proceedings of the Sixth Australia–Japan Joint Workshop on Intelligent and Evolutionary Systems. Australian National University, Canberra, pp 41–46
62. Nehaniv CL (2002) Evolution in asynchronous cellular automata. In: Standish RK, Abbass HA, Bedau MA (eds) Proceedings of the Eighth Conference on Artificial Life. MIT Press, Cambridge, pp 65–74
63. Nakamura K (1974) Asynchronous cellular automata and their computational ability. Syst Comput Controls 5(5):58–66
64. Lee J, Peper F, Adachi S, Morita K (2004) Universal delay-insensitive circuits with bi-directional and buffering lines. IEEE Trans Comput 53(8):1034–1046
65. Adachi S, Peper F, Lee J (2004) Universality of hexagonal asynchronous totalistic cellular automata. In: Chopard B, Hoekstra AG, Sloot PMA (eds) (2004) Cellular Automata, 6th International Conference on Cellular Automata for Research and Industry. LNCS, vol 3305. Springer, Berlin, pp 91–100
66. Takada Y, Isokawa T, Peper F, Matsui N (2006) Construction Universality in purely asynchronous cellular automata. J Comput Syst Sci 72:1368–1385
67. von Neumann J (1966) Theory of Self-Reproducing Automata. Univ Illinois Press, Champaign (edited and completed by Burks AW)
68. Langton CG (1984) Self-reproduction in cellular automata. Phys D 10:135–144
69. Wolfram S (1985) Undecidability and Intractability in Theoretical Physics. Phys Rev Lett 54:735–738
70. Weiss G (2000) Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence. MIT Press, Cambridge

Books and Reviews

References [3, 4, 6, 10, 13, 14, 15, 48, 51] are the main books and reviews. The journal Swarm Intelligence [9] is the main source for new research results. The following provide additional general material related to Swarm Intelligence.

Sahin E, Spears WM (2005) Swarm Robotics: SAB 2004 International Workshop, Santa Monica, CA, July 17, 2004, Revised Selected Papers (Lecture Notes in Computer Science). Springer, Berlin

Dorigo M, Sahin E (2004) Swarm Robotics– special issue editorial. Autonom Robot 17(2–3)111–113

Sipper M (2002) Machine Nature: The Coming Age of Bio-Inspired Computing. McGraw-Hill, New York

Camazine S, Deneubourg J-L, Franks NR, Sneyd J, Theraulaz G, Bonabeau E (2001) Self-Organization in Biological Systems. Princeton Univ Press, Princeton

Symbolic Dynamics

Brian Marcus
Department of Mathematics,
University of British Columbia, Vancouver, Canada

Article Outline

Glossary

In this glossary, we give only brief descriptions of key terms. We refer to specific sections in the text for more precise definitions.

Almost conjugacy (Sect. "Other Coding Problems") A common extension of two shift spaces given by factor codes that are one-to-one almost everywhere.

Automorphism (Sect. "The Conjugacy Problem") An invertible sliding block code from a shift space to itself; equivalently, a shift-commuting homeomorphism from a shift space to itself; equivalently, a topological conjugacy from a shift space to itself.

Dimension group (Sect. "The Conjugacy Problem") A particular group associated to a shift of finite type. This group, together with a distinguished sub-semigroup and an automorphism, captures many invariants of topological conjugacy for shifts of finite type.

Embedding (Sect. "Shift Spaces and Sliding Block Codes") A one-to-one sliding block code from one shift space to another; equivalently, a one-to-one continuous shift-commuting mapping from one shift space to another.

Factor map (Sect. "Shift Spaces and Sliding Block Codes") An onto sliding block code from one shift space to another; equivalently, an onto continuous shift-commuting mapping from one shift space to another. Sometimes called *Factor Code*.

Finite equivalence (Sect. "Other Coding Problems") A common extension of two shift spaces given by finite-to-one factor codes.

Full shift (Sect. "Shift Spaces and Sliding Block Codes") The set of all bi-infinite sequences over an alphabet (together with the shift mapping). Typically, the alphabet is finite.

Higher dimensional shift space (Sect. "Higher Dimensional Shift Spaces") A set of bi-infinite arrays of a given dimension, determined by a collection of finite forbidden arrays. Typically, the alphabet is finite.

Markov partition (Sect. "Origins of Symbolic Dynamics: Modeling of Dynamical Systems") A finite cover of the underlying phase space of a dynamical system, which allows the system to be modeled by a shift of finite type. The elements of the cover are closed sets, which are allowed to intersect only on their boundaries.

Measure of maximal entropy (Sect. "Connections with Information Theory and Ergodic Theory") A shift-invariant measure of maximal measure-theoretic entropy on a shift space. Its measure-theoretic entropy coincides with the topological entropy of the shift space.

Road problem (Sect. "Other Coding Problems") A recently-solved classical problem in symbolic dynamics, graph theory and automata theory.

Run-length limited shift (Sect. "Coding for Data Recording Channels") The set of all bi-infinite binary sequences whose runs of zeros, between two successive ones, are bounded below and above by specific numbers.

Shift equivalence (Sect. "The Conjugacy Problem") An equivalence relation on defining matrices for shifts of finite type. This relation characterizes the corresponding shifts of finite type, up to an eventual notion of topological conjugacy.

Shift space (Sect. "Shift Spaces and Sliding Block Codes") A set of bi-infinite sequences determined by a collection of finite forbidden words; equivalently, a closed shift-invariant subset of a full shift.

Shift of finite type (Sect. "Shifts of Finite Type and Sofic Shifts") A set of bi-infinite sequences determined by a *finite* collection of finite forbidden words.

Sliding block code (Sect. "Shift Spaces and Sliding Block Codes") A mapping from one shift space to another determined by a finite sliding block window; equivalently, a continuous shift-commuting mapping from one shift space to another.

Sofic shift (Sect. "Shifts of Finite Type and Sofic Shifts") A shift space which is a factor of a shift of finite type; equivalently, a set of bi-infinite sequences determined by a finite directed labeled graph.

State splitting (Sect. "The Conjugacy Problem") A splitting of states in a finite directed graph that creates a new graph, whose vertices are the split states. The operation that creates the new graph from the original graph is a basic building block for all topological conjugacies between shifts of finite type.

Strong shift equivalence (Sect. "The Conjugacy Problem") An equivalence relation on defining matrices for shifts of finite type. In principle, this relation characterizes the corresponding shifts of finite type, up to topological conjugacy.

Topological conjugacy (Sect. "Shift Spaces and Sliding Block Codes") A bijective sliding block code from one shift space to another; equivalently, a shift-commuting homeomorphism from one shift space to another. Sometimes called *conjugacy*.

Topological entropy (Sect. "Entropy and Periodic Points") The asymptotic growth rate of the number of finite sequences of given length in a shift space (as the length goes to infinity).

Zeta function (Sect. "Entropy and Periodic Points") An expression for the number of periodic points of each given period in a shift space.

Definition of the Subject

Symbolic dynamics is the study of shift spaces, which consist of infinite or bi-infinite sequences defined by a shift-invariant constraint on the finite-length sub-words. Mappings between two such spaces can be regarded as codes or encodings. Shift spaces are classified, up to various kinds of invertible encodings, by combinatorial, algebraic, topological and measure-theoretic invariants.

The subject is intimately related to many other areas of research, including dynamical systems, ergodic theory, automata theory and information theory. Shift spaces and their associated shift mappings are used to model a rich and important class of smooth dynamical systems and ergodic measure-preserving transformations. These models have provided a valuable tool for classifying and understanding fundamental properties of dynamical systems. In addition, techniques from symbolic dynamics have had profound applications for data recording applications, such as algorithms and analysis of invertible encodings, and problems in matrix theory, such as characterization of the set of eigenvalues of a nonnegative matrix.

Introduction

This article is intended to give a picture of major topics in symbolic dynamics. Section "Origins of Symbolic Dynamics: Modeling of Dynamical Systems" reviews the roots of symbolic dynamics in modeling of dynamical systems. Section "Shift Spaces and Sliding Block Codes" lays the foundation by defining the kinds of spaces and mappings considered in the subject. Section "Shifts of Finite Type and Sofic Shifts" focuses on distinguished special classes of spaces, known as shifts of finite type and sofic shifts. Section "Entropy and Periodic Points" introduces the most fundamental invariants, periodic points and topological entropy. Sections "The Conjugacy Problem" and "Other Coding Problems" survey progress on the conjugacy problem and other classification/coding problems for shifts of finite type and sofic shifts. In Sect. "Coding for Data Recording Channels", we present applications to coding for data recording. Section "Connections with Information Theory and Ergodic Theory" provides a link with information theory and ergodic theory. Finally, Sect. "Higher Dimensional Shift Spaces" treats higher dimensional symbolic dynamics.

While this article covers many of the most important topics in the subject, others have been omitted or treated lightly, due to space limitations. These include one-sided shift spaces, countable state symbolic systems, orbit equivalence, flow equivalence, the automorphism group, cellular automata, and substitution systems. References to work in these sub-areas can be found in the sources mentioned below.

For introductory reading on symbolic dynamics and its applications, beyond this article, one can consult the textbooks Kitchens [65] and Lind and Marcus [78]. There are also excellent introductory survey articles, such as Boyle [23], Lind and Schmidt [79], and S. Williams [132]. In addition there are very good expositions which focus on other aspects of the subject. These include Beal [11], which focuses on connections between symbolic dynamics and automata theory, the lecture notes Marcus–Roth–Siegel [83], which focuses on constrained coding applications, and Immink [54], which focuses on applications to data storage. There are also several excellent collections of articles on special areas of the subject, such as [17,133], and [128]. The book [15] by Berstel and Perrin treats the subject of variable length codes and contains many ideas related to symbolic dynamics. Finally, there is an excellent recent survey by Boyle on Open Problems in Symbolic Dynamics [24].

Origins of Symbolic Dynamics: Modeling of Dynamical Systems

Symbolic dynamics began as an effort to model dynamical systems using sequences of symbols. A *dynamical system* is a pair (X, T) where X is a set and T is a transformation from X to itself. For definiteness, we assume that T is invertible, although this is not a necessary restriction.

Since T maps X to itself, we can iterate T: $T^2 = T \circ T$, $T^3 = T \circ T \circ T$, etc. The *orbit* of a point $x \in X$ is the sequence of points: $\dots, T^{-2}(x), T^{-1}(x), x, T(x), T^2(x), \dots$ In the theory of dynamical systems, one asks questions about orbits such as the following: Are there periodic orbits (i. e., x such that $T^n(x) = x$ for some $n > 0$)? Are there dense orbits (the orbit of x is dense if for any point y in X, $T^n(x)$ is "close" to y for some n)? How does the behavior of an orbit vary with x? How can we describe the collection of all orbits of the dynamical system? When is the dynamical system "chaotic"? For more information on dynamical systems, we refer the reader to [16,38,47].

The subject of dynamical systems has its roots in Classical Mechanics; in that setting, X is the set of all possible states of a system (e. g., the positions, momenta of all particles in a physical system), and the transformation T is the time evolution map, which maps the state of the system at one time to the state of the system at one time unit later.

Symbolic dynamics provides a model for the orbits of a dynamical system (X, T) via a space of sequences. This is done by "quantizing" X into cells, associating symbols to the cells and representing points as bi-infinite sequences of symbols. For instance in Fig. 1, X is a square, and T is some transformation of the square.

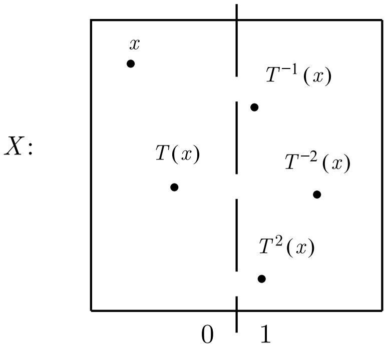
We have drawn a portion of the orbit of a point $x \in X$ for the dynamical system (X, T). We have also quantized X into two cells: the left half, called '0', and the right half, called '1'. Then the point x is represented by the bi-infinite sequence $s(x) = \dots s_{-2}s_{-1}.s_0s_1s_2 \dots$ where s_n is the label of the cell to which $T^n(x)$ belongs (here, we use the decimal point to separate coordinates s_i, $i < 0$ from s_i, $i \geq 0$). So, for x as given in Fig. 1, we see that

$$s(x) = \dots 11.001 \dots$$

for instance $s_0 = 0$ because x belongs to the left half of the square, and $s_2 = 1$ because $T^2(x)$ belongs to the right half of the square. Now, if x is represented by the sequence $s(x)$, then $T(x)$ is represented by the shift of $s(x)$:

$$s(T(x)) = \dots 110.01 \dots .$$

So, T is represented 'symbolically' as the shift transformation.



Symbolic Dynamics, Figure 1
Representing points symbolically

By representing all points of X as bi-infinite sequences, we obtain a *symbolic dynamical system* (Y, σ) where Y is a set of sequences (representing X) and σ is the shift transformation (representing T). The "symbolic" refers to the symbols, and the "dynamical" refers to the action of the shift transformation.

For this representation to be faithful, distinct points should be represented by distinct sequences, and this imposes extra conditions on how X is quantized into cells. Also, Y is typically a set of sequences constrained by certain rules, such as a certain symbol may only be followed by certain other symbols.

In this way, one can use symbolic dynamics to study dynamical systems. Properties of orbits of the original dynamical system are reflected in properties of the resulting sequences. For instance, a point whose orbit is periodic becomes a periodic sequence, and the distribution of the orbit of a point x in X is reflected in the distribution of finite strings within $s(x)$. Beginning with Hadamard [46] in 1898 and followed by Hedlund, Morse and others in the 1920s, 1930s and 1940s [48,49,92,93], this method was used to prove the existence of periodic, almost periodic and other interesting motions in classical dynamical systems, such as geodesic flows on surfaces of negative curvature; this was done by finding interesting sequences satisfying the constraints defined by the corresponding symbolic dynamical system. Later on, this was extended to general hyperbolic systems, where the symbolic dynamics is constructed using a *Markov Partition*, which is a disjoint collection of open sets whose closures cover X, each of which looks like a "rectangle", with vertical (resp., horizontal) fibers contracted (resp., expanded) by T. Markov Partitions were developed by Adler and Weiss [5], Sinai [120] and Bowen [18,19]; see also [13] and Sect. 6.5 in [78].

In more recent years, symbolic dynamics has been used as a tool in classification problems for dynamical systems. Here, the problem of determining when one dynamical system is 'equivalent' to another becomes, via symbolic dynamics, a coding problem. Roughly speaking, two dynamical systems, (X_1, T_1) and (X_2, T_2), are *equivalent* if there is an invertible mapping from X_1 to X_2 which makes T_1 "look like" T_2. If the corresponding symbolic dynamical systems are denoted (Y_1, σ) and (Y_2, σ), then an equivalence between (X_1, T_1) and (X_2, T_2) becomes a time-invariant, invertible encoding from Y_1 to Y_2 (time-invariant because the shift transformation represents the dynamics). Thus, the classification problem in dynamical systems leads to a coding problem between constrained sets of sequences.

We have described dynamical systems as the discrete-time iteration of a single mapping. However, continuous-time iterations have been studied since the inception of the subject. These are known as continuous-time flows, with the main example being the set of solutions to a system of ordinary differential equations. Indeed, the work of Hedlund and Morse mentioned above was done in this context.

Shift Spaces and Sliding Block Codes

Let $\mathcal{A}$ be an alphabet of symbols, which we assume to be finite. The principal objects of study in symbolic dynamics are certain kinds of collections of sequences of symbols from $\mathcal{A}$. Typically, these sequences are infinite $x = x_0x_1x_2\ldots$, but it is often more convenient to deal with *bi-infinite sequences* $x = \ldots x_{-2}x_{-1}x_0x_1x_2\ldots$ For some problems, the results are similar in the infinite and bi-infinite categories, while for other problems, they are quite different. In this article, we focus on the bi-infinite setting.

The symbol x_i is the ith *coordinate* of x. When writing a specific sequence, we need to specify which is the 0th coordinate. As suggested in Sect. "Origins of Symbolic Dynamics: Modeling of Dynamical Systems", this is done with a decimal point to separate the x_i with $i \geq 0$ from those with $i < 0$: $x = \ldots x_{-2}x_{-1}.x_0x_1x_2\ldots$. A *block* or *word* over $\mathcal{A}$ is a finite sequence of symbols from $\mathcal{A}$. A block of length N is called an N-*block*. For blocks u, v, the block uv is the concatenation of u and v, and for a block w, the concatenation of N copies of w is denoted w^N.

The *full $\mathcal{A}$-shift* $\mathcal{A}^{\mathbf{Z}}$ is the set of all bi-infinite sequences of symbols from $\mathcal{A}$. The *full r-shift* is the full shift over the alphabet $\{0, 1, \ldots, r-1\}$. The *shift map* σ on a full shift maps a point x to the point $y = \sigma(x)$ whose ith coordinate is $y_i = x_{i+1}$.

The *orbit* of a point in a full shift is its orbit under the shift map. The full shift contains many different types of orbits. For instance, it contains a dense orbit (namely, any sequence which contains every block in the alphabet) and periodic orbits (namely, any sequence which is periodic).

We are interested in sets that can be specified by a list (finite or infinite) of forbidden blocks. Namely, given a collection $\mathcal{F}$ of "forbidden blocks" over $\mathcal{A}$, the subset X consisting of all sequences in $\mathcal{A}^{\mathbf{Z}}$, none of whose subwords belong to $\mathcal{F}$, is called a *shift space* (or simply *shift*), and we write $X = X_{\mathcal{F}}$. When a shift space X is contained in a shift space Y, we say that X is a *subshift* of Y.

Example 1 X is the set of all binary sequences with no two 1's next to each other. Here $X = X_{\mathcal{F}}$, where $\mathcal{F} = \{11\}$. This shift is called the *golden mean shift*, for reasons that will become apparent later.

Example 2 X is the set of all binary sequences so that between any two 1's there are an even number of 0's. We can take for $\mathcal{F}$ the collection

$$\{10^{2n+1}1 \colon n \geq 0\}\,.$$

This example is naturally called the *even shift*.

Example 3 X is the set of all binary sequences such that between any two successive 1's number of 0's is prime. We can take for $\mathcal{F}$ the collection

$$\{10^n1 \colon n \text{ is composite}\}\,.$$

This example is naturally called the *prime shift*.

Alternatively (and equivalently), shift spaces can be defined as closed, shift-invariant subsets of full shifts. Here, "closed" means with respect to a metric, for which two points are close if they agree in a large "central block"; one such metric is $\rho(x, y) = 2^{-k}$ if $x \neq y$, with k maximal such that $x_{[-k,k]} = y_{[-k,k]}$ (with the conventions that $\rho(x, y) = 0$ if $x = y$ and $\rho(x, y) = 2$ if $x_0 \neq y_0$).

Let X be a subset of a full shift, and let $\mathcal{B}_N(X)$ denote the set of all N-blocks that occur in elements of X. The *language of* X is $\mathcal{B}(X) = \cup_N \mathcal{B}_N(X)$. It can be shown that the language of a shift space determines the shift space uniquely, and so we can equally well describe a shift space by specifying the "occurring" or "allowed" blocks, rather than the forbidden blocks. For example, the golden mean shift is specified by the language of blocks in which 1's are isolated.

This establishes a connection with automata theory [6,11], which studies collections of blocks, rather than infinite or bi-infinite sequences. The languages that occur in symbolic dynamics, i. e. as $\mathcal{B}(X)$ for some shift space X, are simply those sets $\mathcal{L}$ of blocks that satisfy two simple properties: every sub-block of an element of $\mathcal{L}$ belongs to $\mathcal{L}$, and every element $w \in \mathcal{L}$ is extendable to a larger block $awb \in \mathcal{L}$, with $a, b \in \mathcal{A}$.

Some examples are best described by allowed blocks. One example is the famous *Morse shift*.

Example 4 Let $A_0 = 0$ and inductively define blocks $A_{n+1} = A_n\overline{A_n}$, where $\overline{A_n}$ denotes bitwise complement. The shift space whose allowed blocks are the sublocks of the A_n is called the *Morse shift*.

This shift space has an alternative description as follows. Since each A_n is a prefix of A_{n+1}, the sequence of blocks A_n determines a unique right-infinite sequence x^+ (with each A_n as a a prefix). If we denote x^- as the left-infinite sequence obtained by writing x^+ backwards, then the Morse shift is the closure of the orbit of the point $x^-.x^+$. The sequence x^+ is known as the Prouhet–Thue–Morse (PTM) sequence, since Prouhet and Thue introduced it earlier, but for different purposes.

While it is not immediately obvious, it can be shown that the PTM sequence is not periodic; moreover, the Morse shift is a *minimal shift*, which means that all its orbits are dense [91]. In contrast, while the full, golden mean, even and prime shifts all have dense orbits, each also has a dense set of periodic orbits.

There are two simple, but very important, constructions in symbolic dynamics that construct from a given shift space a new version which in some sense looks deeper into the space at the cost of a larger alphabet and more complex description.

Let X be a shift space over the alphabet $\mathcal{A}$, and $\mathcal{A}^{(N)} = \mathcal{B}_N(X)$. We can consider $\mathcal{A}^{(N)}$ as an alphabet in its own right, and form the full shift $(\mathcal{A}^{(N)})^{\mathbf{Z}}$. Define the *Nth higher block code* by

$$(\beta_N(x))_i = x_{[i,i+N-1]} \, .$$

Then the Nth *higher block shift* or *higher block presentation* of a shift space X is the image $X^{[N]} = \beta_N(X)$ in the full shift over $\mathcal{A}^{(N)}$.

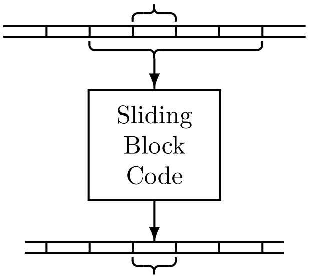
Similarly, define the *Nth higher power code* $\gamma_N \colon X \to (\mathcal{A}^{(N)})^{\mathbf{Z}}$ by

$$(\gamma_N(x))_i = x_{[iN,iN+N-1]} \, .$$

The Nth *higher power shift* X^N of X is the image $X^N = \gamma_N(X)$ of X. The difference between $X^{[N]}$ and X^N is that the former is constructed by considering overlapping blocks and the latter by non-overlapping blocks.

Next, we turn to mappings between shift spaces. Suppose that $x = \ldots x_{-1}.x_0x_1\ldots$ is a sequence in a shift space X over $\mathcal{A}$. We can transform x into a new sequence $y = \ldots y_{-1}.y_0y_1\ldots$ over another alphabet C as follows. Fix integers m and n with $-m \le n$. To compute the ith coordinate y_i of the transformed sequence, we use a function Φ that depends on the "window" of coordinates of x from position $i-m$ to position $i+n$. Here $\Phi\colon \mathcal{B}_{m+n+1}(X) \to C$ is a fixed *block map*, called an $(m+n+1)$-*block map* from allowed $(m+n+1)$-blocks in X to symbols in C, and so

$$y_i = \Phi(x_{i-m}x_{i-m+1}\ldots x_{i+n}) = \Phi(x_{[i-m,i+n]}) \, .$$



Symbolic Dynamics, Figure 2
Sliding block code

This is illustrated in Fig. 2.

Let X be a shift space over $\mathcal{A}$, and $\Phi\colon \mathcal{B}_{m+n+1}(X) \to C$ be a block map. Then the map $\phi\colon X \to C^{\mathbf{Z}}$ defined by $y = \phi(x)$, with y_i given by Φ above, is called the *sliding block code* with *memory m* and *anticipation n induced by Φ*. We will denote the formation of ϕ from Φ by $\phi = \Phi_\infty^{[-m,n]}$, or more simply by $\phi = \Phi_\infty$ if the memory and anticipation of ϕ are understood. If not specified, the memory is taken to be 0. If Y is a shift space contained in $C^{\mathbf{Z}}$ and $\phi(X) \subseteq Y$, we write $\phi\colon X \to Y$.

In analogy with the characterization of shift spaces as closed shift-invariant sets, sliding block codes can be characterized in a topological manner: namely, as the maps between shift spaces that are continuous and commute with the shift. This result is known as the Curtis–Hedlund–Lyndon theorem [50].

Example 5 Let $\mathcal{A} = \{0,1\} = C$, $X = \mathcal{A}^{\mathbf{Z}}$, $m = 0$, $n = 1$, and $\Phi(a_0a_1) = a_0 + a_1 \pmod 2$. Let $\phi = \Phi_\infty\colon X \to X$.

Example 6 The sliding block code, generated by $\Phi(00) = 1$, $\Phi(01) = 0 = \Phi(10)$, maps the golden mean shift onto the even shift.

Example 7 There is a trivial sliding block code from the full 2-shift into the full 3-shift, generated by $\Phi(0) = 0$, $\Phi(1) = 1$.

If a sliding block code $\phi\colon X \to Y$ is onto, then ϕ is called a *factor code* or *factor map*, and Y is a *factor* of X. If $\phi\colon X \to Y$ is one-to-one, then ϕ is called an *embedding* of X into Y. The sliding block code in Example 7 is an embedding but not a factor code, while the codes in Examples 5 and 6 are factor maps, but not embeddings.

A major (and unrealistic) goal of symbolic dynamics is to classify in an explicit way shift spaces up to the following natural notion of equivalence. A sliding block code $\phi\colon X \to Y$ is a *conjugacy* (or *topological conjugacy*) if it is invertible with sliding block inverse. Equivalently, a conjugacy is a bijective sliding block code and therefore simultaneously a factor code and an embedding. If there is a conjugacy from one shift space X to another Y, we say that X and Y are *conjugate*, denoted $X \cong Y$.

As an example, the higher block map β_N is a conjugacy between a shift space X and its higher block shift $X^{[N]}$. Via this code, we can "re-code" any sliding block code as a 1-block code (though typically a conjugacy and its inverse cannot, by this artifice, be simultaneously re-coded to 1-block codes).

In this section, we have given examples of relatively simple sliding block codes. But the typical conjugacy, as well as factor code and embedding, can be much more complicated.

Shifts of Finite Type and Sofic Shifts

A *shift of finite type (SFT)* is a shift space that can be described by a finite set of forbidden blocks, i. e., a shift space X having the form $X_{\mathcal{F}}$ for some finite set $\mathcal{F}$ of blocks. The terminology shift of finite type (or *subshift of finite type*) comes from dynamical systems (Smale [121]).

An SFT is *M-step* (or has *memory M*) if it can be described by a collection of forbidden blocks all of which have length $M+1$. It is easy to see that any SFT is M-step for some M.

Since any shift space can be defined by many different collections of forbidden blocks, it is useful to have the following equivalent condition expressed in terms of allowed blocks: an SFT is M-step if and only if whenever u is an allowed block of length at least M, u' is the suffix of u with length M and a is a symbol, then ua is allowed if and only if $u'a$ is allowed. In other words, in order to tell whether a symbol can be allowably concatenated to the end of an allowed word u, one need only look at the last M symbols of u. This is analogous to the "finite memory" property of M-step Markov chains.

The golden mean shift X is is a 1-step SFT, since it was defined by a forbidden list consisting of exactly one block: $\mathcal{F} = \{11\}$. Equivalently, it is only the last symbol of an allowed block that determines whether a given symbol can be concatenated at the end. In contrast, the even shift is not an SFT: for any M, the symbol 1 can be concatenated to the end of exactly one of the (allowed) words 10^M and 10^{M+1}.

Recall that the higher block code β_M is a conjugacy from X to $X^{[M]}$. Via this code, any SFT can be recoded to a 1-step SFT. And so any sliding block code on a shift space can be recoded to a 1-block code on a 1-step SFT. It is useful to have a concrete description of 1-step SFT's. In fact, these are precisely the shift spaces consisting of all bi-infinite sequences of vertices along paths on a finite directed graph. These are called *vertex shifts*. We find it more convenient to work with sequences of edges instead. To be precise:

Let G be a finite directed graph (or simply graph) with vertices (or *states*) $\mathcal{V} = \mathcal{V}(G)$ and *edges* $\mathcal{E} = \mathcal{E}(G)$. For an edge e, $i(e)$ denotes the initial state and $t(e)$ the terminal state. A *path* in G is a finite sequence of edges in G such that the terminal state of an edge coincides with the initial state of the following edge; a *cycle* in G is path that begins and ends at the same state. We will assume that G is *essential*, i. e., that every state has at least one outgoing edge and one incoming edge.

The *adjacency matrix* $A = A(G)$ is the matrix indexed by $\mathcal{V}$ with A_{IJ} equal to the the number of edges in G with initial state I and terminal state J. Since a graph and its adjacency matrix essentially determine the same information, we will frequently associate a graph G with its adjacency matrix A and a nonnegative integer matrix A with a graph G.

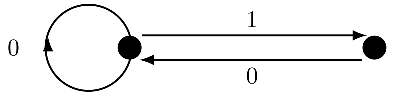
The *edge shift* X_G or X_A is the shift space over the alphabet $\mathcal{A} = \mathcal{E}$ defined by

$$\begin{aligned} X_G &= X_A \\ &= \left\{\xi = (\xi_i)_{i\in\mathbf{Z}} \in \mathcal{E}^{\mathbf{Z}} \colon \text{ each } \xi_{i+1} \text{ follows } \xi_i\right\} . \end{aligned}$$

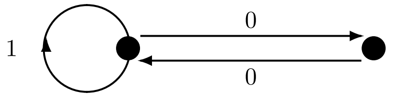
It can be readily verified that edge shifts are 1-step SFT's. While edge shifts do not include all 1-step SFT's, any 1-step SFT can be recoded to an edge shift, and, compared with vertex shifts, edge shifts offer the advantage of a more compact description.

For many purposes, one can study a general shift space X by breaking it into smaller, more well-behaved pieces. A shift space is *irreducible* if whenever u and w are allowed blocks, there is a "connecting" block v such that uvw is allowed. While shift spaces do not always decompose into disjoint unions of irreducible shifts, every SFT can be written as a finite disjoint union of irreducible SFT's X_i together with "transient" one-way connections from one X_i to another. And irreducible edge shifts can be characterized in a particularly concrete form: namely, X_G is irreducible if and only if G is *irreducible*, i. e., for every ordered pair of vertices I and J there is a path in G starting at I and ending at J.

There is a stronger notion which is defined by a uniformity condition on the length of the connecting block. A shift space is *mixing* if whenever u and w are allowed blocks, there is an N, possibly depending on u and w, such that for all $n \geq N$, there is block v of length n such that uvw is allowed. And an edge shift X_G is mixing if and only if G is *primitive*, i. e., there is an integer N such that for any $n \geq N$ and any ordered pair of vertices I and J there is



Symbolic Dynamics, Figure 3
Presentation of golden mean shift



Symbolic Dynamics, Figure 4
Presentation of even shift

a path in G of length n starting at I and terminating at J. It follows that for SFT's in the definition of mixing, the uniform connecting length N can be chosen independent of the allowed blocks u and w.

It can be shown that, in some sense, any irreducible SFT X can be broken down into a union of disjoint maximal mixing shifts; namely, X can be written as the disjoint union of finitely many sets X_i, $i = 0, \ldots, p-1$ such that $\sigma(X_i) = X_{i+1} \bmod p$ and for each i, σ^p restricted to X_i can be regarded as a mixing SFT. This is a consequence of Perron–Frobenius theory, upon which symbolic dynamics relies heavily; see Seneta [118] for an introduction to this theory.

A *sofic shift* is the set of bi–infinite sequences obtained from a finite labeled directed graph $\mathcal{G} = (G, \mathcal{L})$; here, G is a finite directed graph and $\mathcal{L}$ is a labeling of the edges of G. The labeled graph is often called a *presentation* of the sofic shift. The golden mean shift and even shift are sofic, with presentations given in Figs. 3 and 4.

SFT's are sofic because any M-step SFT can be presented by a graph whose states are allowed M-blocks. Note also that any sofic shift is a factor of an SFT, namely via a (1-block) factor code $\mathcal{L}_\infty$ on the edge shift X_G based on a presentation $(G, \mathcal{L})$. In fact, the converse is true, and so the sofic shifts are precisely the shift spaces that are factors of SFT's. This was the original definition of sofic shifts given by Weiss [130].

Typically, the labeling is *right resolving*, which means that at any given state, all outgoing edges have distinct labels (as in Figs. 3 and 4).

Theorem 1

(a) *Any sofic shift has a right resolving presentation.*
(b) *Any irreducible sofic shift has a unique minimal right resolving presentation.*

Part (a) is a direct consequence of the subset construction in automata theory [6,11] which constructs a right resolving presentation from an arbitrary presentation; see also Coven and Paul [33,34]. Part (b) makes use of the the state-minimization algorithm from automata theory [6,11], but requires an idea beyond that found in automata theory (Fischer [39,40]). The unique presentation in part (b) may be regarded as a canonical presentation. We remark that for an irreducible (resp. mixing) sofic shift, the underlying graph of the unique minimal right resolving presentation is irreducible (resp., primitive).

Sometimes it is useful to weaken the concept of right resolving to *right closing*, which means "right resolving with delay"; more precisely a labeling is right closing, with delay D if all paths of length $D+1$ with the same initial state and the same label have the same initial edge. Also, we sometimes consider *left resolving* and *left closing* labellings (replace "outgoing" with "incoming" in the definition of right resolving, and replace "initial" with "terminal" in the definition of right closing).

While the class of SFT's is defined by a "finite-memory" property, the more general class of sofic shifts is defined by a "finite-state" property, in that the possible symbols that can occur at time 0 are determined by the past via one of finitely many states. In a presentation, the vertices can be viewed as state information which connects sequences in the past with sequences in the future.

It is not difficult to show that neither the prime shift (Example 3) nor the Morse shift (Example 4) are sofic and therefore also not SFT. For the prime shift, this can be done by an application of the pumping lemma from automata theory [6] (or p. 68 of [78]).

There are uncountably many shift spaces, but only countably many sofic shifts. So, it is not surprising that the behavior of sofic shifts is very special. However, they are very useful in modeling smooth dynamical systems (Sect. "Origins of Symbolic Dynamics: Modeling of Dynamical Systems") and in information theory and applications to data recording (Sect. "Coding for Data Recording Channels"), where they arise as *constrained systems*, although this term is usually reserved for the set of (finite) blocks obtained from a finite directed labeled graph [83].

Entropy and Periodic Points

An invariant of conjugacy is an object associated to a shift space that is preserved under conjugacy. It can be shown that many of the concepts that we have already introduced are invariants: irreducibility, mixing, as well as the properties of being a shift of finite type or sofic shift. Beyond these qualitative invariants, there are many quantitative invariants that can, in many cases, be computed explicitly. Foremost among these is topological entropy.

The *(topological) entropy* (or simply *entropy*) of X is:

$$h(X) = \lim_{N \to \infty} \frac{\log |\mathcal{B}_N(X)|}{N} ;$$

here, $|\cdot|$ denotes cardinality, and for definiteness the log is taken to mean $\log_2$ but any base will do. A subadditivity argument shows that the limit does indeed exist [127]. It should be evident that $h(X)$ is a measure of the "size" or "complexity" of X, as it is simply the asymptotic growth rate of the number of blocks that occur in X. Topological entropy for continuous dynamical systems was defined by Adler, Konheim and McAndrew [3], in analogy with measure-theoretic entropy.

Since a k-block sliding block code from a shift space X to a shift space Y maps $\mathcal{B}_{N+k-1}(X)$ into $\mathcal{B}_N(Y)$, it is not hard to see that entropy is an invariant of conjugacy and that it cannot increase under factors and cannot decrease under embeddings.

In many cases, one can explicitly compute entropy. For example, for the full r-shift X, $|\mathcal{B}_N(X)| = r^N$, and so $h(X) = \log r$. And from the defining sequence of the Morse shift, we see that the number of distinct 2^N-blocks is at most $4(2^N)$, it follows that the growth rate cannot be exponential and so the entropy of the Morse shift is zero.

For SFT's and more generally for sofic shifts, entropy can be computed explicitly. The key to this computation is the following result, which is based on the Perron–Frobenius theorem.

Theorem 2 [99,119] *For any graph G, $h(X_G) = \log \lambda_{A(G)}$, where $\lambda_{A(G)}$ is the largest eigenvalue of $A(G)$.*

The rough idea is that the number of N-blocks in X_G is the number of paths of length N and thus also the sum of the entries of $A(G)^N$, which is controlled by the largest eigenvalue. This is proven first for primitive graphs, whose adjacency matrices have a unique eigenvalue of maximum modulus and this eigenvalue is positive; in fact, for a primitive graph and a pair of states I, J, the number of paths of length N from I to J grows like $\lambda_{A(G)}^N = 2^{Nh(X_G)}$. For general graphs, one uses the decomposition into primitive, and then irreducible, graphs.

To extend this to a sofic shift Y, one uses a right resolving presentation $(G, \mathcal{L})$ of Y. Since every block of Y is the label of at most most $\mathcal{V}(G)$ paths in G, it follows that:

Theorem 3 *Let $\mathcal{G} = (G, \mathcal{L})$ be a right-resolving labeled graph presenting a sofic shift Y. Then $h(Y) = h(X_G)$.*

From Fig. 3, we see that the golden mean shift is obtained as a right resolving presentation of the graph with adjacency matrix:

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} .$$

A computation shows that λ_A is the golden mean, and so the entropy of the golden mean shift is the log of the golden mean; this is one explanation of the meaning of the term golden mean shift. From Fig. 4, we see that the even shift has the same entropy.

One cannot overstate the importance of entropy as an invariant. Yet, it is somewhat crude; it is perhaps not surprising that a single numerical invariant would not be sufficient to completely capture the many intricacies of shift spaces, even of sofic shifts or SFT's.

An invariant finer than entropy is the zeta function, which combines information regarding the numbers of periodic sequences of all periods, described as follows.

For a shift space X, let $p_n(X)$ denote the number of points in X of period n (i. e., the number of $x \in X$ such that $\sigma^n(x) = x$). It is straightforward to show that each $p_n(X)$ is an invariant. Since distinct periodic sequences define distinct blocks, it follows that

$$\limsup_{n \to \infty} \frac{1}{n} \log p_n(X) \le h(X) .$$

The inequality can be strict; for example, there are shift spaces (such as the direct product of the Morse shift with the full 2-shift) with positive entropy but no periodic points at all.

However, for irreducible SFT's and sofic shifts, the entropy $h(X)$ can be recovered from the sequence $p_n(X)$. The key to understanding this is the fact that for an edge shift $X = X_A$, $p_n(X) = \mathrm{tr}(A^n)$ and thus is the sum of the nth powers of the (non-zero) eigenvalues of A; in the case that A is primitive, the largest eigenvalue λ_A strictly dominates the other eigenvalues, and thus for large n,

$$\log p_n(X) = \log \mathrm{tr}(A^n) \sim n \log \lambda_A \sim nh(X) .$$

This shows that for a mixing SFT, the entropy equals the growth rate of numbers of periodic points. In fact, this result applies to all SFT's and sofic shifts.

Theorem 4 *For a sofic shift X,*

$$\limsup_{n \to \infty} \frac{1}{n} \log p_n(X) = h(X) .$$

The lim sup turns out be a limit in the case that X is a mixing sofic shift.

Most of what we have stated for $p_n(X)$ here applies equally well to $q_n(X)$, the number of points of *least* period n in X. This follows from the fact that "most" periodic points of period n have least period n.

The periodic point information can be conveniently combined into a single invariant, known as the *zeta function*. For a shift space X,

$$\zeta_X(t) = \exp\left(\sum_{n=1}^{\infty} \frac{p_n(X)}{n} t^n\right) .$$

For an edge shift X_A, one computes the zeta function to be the reciprocal of a polynomial:

$$\zeta_{X_A}(t) = \frac{1}{t^r\,\chi_A(t^{-1})} = \frac{1}{\det(I - tA)} ,$$

which is completely determined by the non-zero eigenvalues (with multiplicity) of A (Bowen and Lanford [21]). As an example, the zeta function of the golden mean shift is $1/(1-t-t^2)$.

The discussion above applies equally well to SFT's since they are conjugate to edge shifts. For a sofic shift, the zeta function turns out to be a rational function, i. e., quotient of two polynomials. This can be shown by analyzing properties of a right resolving presentation of the sofic shift. From this it turns out that the zeta function of the even shift is $(1-t)/(1-t-t^2)$. The technique for computing zeta functions of sofic shifts was developed by Manning [80] (actually, Manning developed the technique to compute zeta functions of hyperbolic dynamical systems).

So, for sofic shifts, all of the periodic point information is determined by a finite collection of complex numbers, namely the zeros and poles of the zeta function.

Finally, we mention another simple invariant obtained from the periodic points. The *period*, per(X), of a shift space X is the gcd of lengths of periodic points in X, i. e., the gcd of the set of n such that $p_n(X) \neq 0$. If $X = X_G$ is an edge shift and G is irreducible, then $\mathrm{per}(X_G) = \mathrm{per}(G)$, which is defined to be the gcd of cycle lengths in G and coincides with the gcd of the lengths of all cycles in G based at any given state. For an irreducible graph with period p and any state I, the number of cycles of length N, a multiple of p, at I grows like $\lambda^N_{A(G)} = 2^{Nh(X_G)}$. Also, an irreducible graph G is primitive if $\mathrm{per}(G) = 1$.

The Conjugacy Problem

The conjugacy problem for SFT's and sofic shifts is a major open problem. After much effort, it remains unsolved today. Much of what we know goes back to R. Williams [131] in the 1970s. One of Williams' main results was that any conjugacy can be decomposed into simple building blocks, as follows.

Let A and B be nonnegative integral matrices, with associated graphs G and H. An *elementary equivalence from A to B* is a pair (R, S) of rectangular nonnegative integral matrices satisfying

$$A = RS\,, \quad B = SR\,. \tag{1}$$

In this case we write $(R, S)\colon A \equiv B$. A *strong shift equivalence of lag ℓ from A to B* is a sequence of ℓ elementary equivalences

$$\begin{aligned}&(R_1, S_1)\colon A = A_0 \equiv A_1\,,\\ &(R_2, S_2)\colon A_1 \equiv A_2\,,\\ &\ldots\,,\\ &(R_\ell, S_\ell)\colon A_{\ell-1} \equiv A_\ell = B\,.\end{aligned}$$

In this case we write $A \approx B(\text{lag } \ell)$. We say that A is *strong shift equivalent to B* (and write $A \approx B$) if there is a strong shift equivalence of some lag from A to B.

Via the matrix Equation (1), one creates a graph K whose state set is the disjoint union of $\mathcal{V}(G)$ and $\mathcal{V}(H)$; for each $I \in V_G$ and $J \in V_H$, the graph has R_{IJ} edges from I to J and S_{JI} edges from J to I. The equation $A = RS$ allows one to associate each edge e of G with a unique path $r(e)s(e)$ of length two running from $\mathcal{V}(G)$ to $\mathcal{V}(H)$ to $\mathcal{V}(G)$. Similarly, the equation $B = SR$ allows one to associate each edge e of H with a path $s(e)r(e)$ of length two running from $\mathcal{V}(H)$ to $\mathcal{V}(G)$ to $\mathcal{V}(H)$. One can show that the 2-block sliding block code defined by $\Phi(ef) = s(e)\,r(f)$ defines a conjugacy from X_A to X_B, and so whenever $A \approx B$, $X_A \cong X_B$. Williams proved this and its converse:

Theorem 5 *(R. Williams [131]) The edge shifts X_A and X_B are conjugate if and only if A and B are strong shift equivalent.*

In fact, Williams showed that any conjugacy can be decomposed into a composition of conjugacies defined by elementary equivalences. This can be interpreted in a way that shows that X_A and X_B are conjugate if and only if we can pass from G to H by a sequence of state splittings and amalgamations, defined as follows.

Let G be a graph with states $\mathcal{V}$ and edges $\mathcal{E}$. For each $I \in \mathcal{V}$, partition the outgoing edges from I into disjoint nonempty sets $\mathcal{E}^1_I, \mathcal{E}^2_I, \ldots, \mathcal{E}^{m(I)}_I$. Let $\mathcal{P}$ denote the resulting partition of $\mathcal{E}$, and let $\mathcal{P}_I$ denote the partition $\mathcal{P}$ restricted to $\mathcal{E}_I$. The *out-split graph H formed from G using $\mathcal{P}$* has states $I^1, I^2, \ldots, I^{m(I)}$, where I ranges over the states in $\mathcal{V}$, and edges e^j, where e is any edge in $\mathcal{E}$ and

$1 \le j \le m(t(e))$. If $e \in \mathcal{E}$ goes from I to J, then $e \in \mathcal{E}_I^i$ for some i, and we define the e^j to have initial state I^i and terminal state J^j, The 2-block code generated by $\Phi(ef) = e^j$, where $f \in \mathcal{E}_{t(e)}^j$, defines a conjugacy, called an out-splitting code, from X_G to X_H. Similarly, one defines an in-splitting code. Inverses of these conjugacies are called out-amalgamation and in-amalgamation codes.

Figure 5 depicts an out-splitting. The graph (a) on the left has three states I, J, K and the partition elements that define the splitting are $\mathcal{E}_I^1 = \{a\}, \mathcal{E}_I^2 = \{b, c\}, \mathcal{E}_J^1 = \{d\}$, $\mathcal{E}_K^1 = \{e\}, \mathcal{E}_K^2 = \{f\}$; the graph (b) is the resulting split graph.

By interpreting state splitting and amalgamation in terms of adjacency matrices one can show that such operations generate elementary equivalences. And one can decompose elementary equivalences into splittings and amalgamations. It follows that:

Theorem 6 (R. Williams [131]) *Every conjugacy from one edge shift to another is the composition of splitting codes and amalgamation codes.*

This classification for edge shifts naturally extends to SFT's since every SFT is conjugate to an edge shift. It also extends to sofic shifts, and we describe this in the context of irreducible sofic shifts.

Recall that an irreducible sofic shift has a unique minimal right resolving presentation. Any labeled graph can be completely described by a symbolic adjacency matrix, which records the transitions (edges) in the underlying physical graph, as well as the labels of the edges. Namely, the symbolic adjacency matrix is indexed by the states of the underlying graph and the (I, J)-entry is the formal sum of the labels of edges from I to J. It turns out that the notions of elementary equivalence, and hence strong shift equivalence, can be extended to more general categories, in particular to symbolic adjacency matrices.

Theorem 7 (Krieger [72], Nasu [97]) *Let X and Y be irreducible sofic shifts. Let A and B be the symbolic adjacency matrices of the minimal right-resolving presentations of X and Y, respectively. Then X and Y are conjugate if and only if A and B are strong shift equivalent.*

The classification, provided by these results, would be of limited use if the story ended here. Fortunately, Williams showed that strong shift equivalence yields a strong, delicate and somewhat computable necessary condition for conjugacy.

Let A and B be nonnegative integral matrices and $\ell \ge 1$. A *shift equivalence* of lag ℓ is a pair (R, S) of rectangular nonnegative integral matrices satisfying the shift equivalence equations

$$AR = RB\,, \quad SA = BS\,, \quad A^\ell = RS\,, \quad B^\ell = SR\,.$$

We denote this situation by $(R, S)\colon A \sim B(\text{lag } l)$. We say that A is *shift equivalent* to B, written $A \sim B$, if there is a shift equivalence from A to B of some lag. It is not hard to see that an elementary equivalence is a shift equivalence of lag 1 and that shift equivalence is an equivalence relation. It follows that:

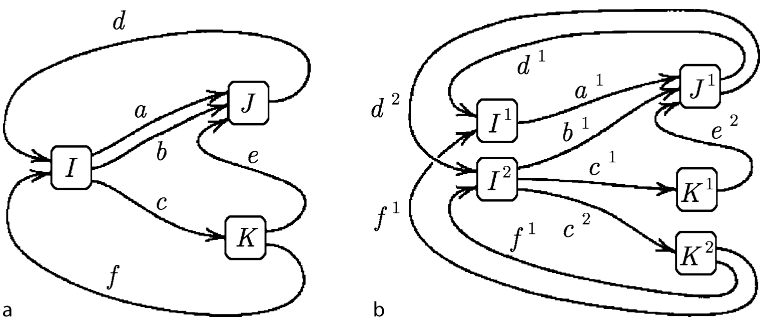
Theorem 8 (Williams [131]) *Strong shift equivalence implies shift equivalence. More precisely, if $A \approx B(\text{lag } \ell)$, then $A \sim B(\text{lag } \ell)$.*

Recall from Sect. "Entropy and Periodic Points" that for an edge shift X_A, the set of nonzero eigenvalues, with multiplicity, of A determines the zeta function and hence this set is an invariant of conjugacy. Using the shift equivalence equations, one can show more: the entire Jordan form corresponding to the nonzero eigenvalues is an invariant. This information depends only on properties of the adjacency matrix considered as a linear transformation (over $\mathbf{R}$ or $\mathbf{Q}$). However, A is a nonnegative, integral matrix and both nonnegativity and integrality provide substantially more information. One such invariant, that follows from shift equivalence and makes use of integrality is the Bowen–Franks group [20], $BF(A) = Z^r/Z^r(I - A)$.

Until recently all of the information contained in known conjugacy invariants, such as those above, was subsumed in shift equivalence. And Kim and Roush showed that shift equivalence is decidable [60,61], meaning that there is a finite decision procedure via a Turing machine that decides whether two given edge shifts, and therefore two given SFT's, are conjugate (they also showed that a notion of shift equivalence for sofic shifts, formulated by Boyle and Krieger [26], is decidable [62]). So, a central focus of the subject was the question: is shift equivalence a complete invariant of conjugacy? The answer turns out to be No, as proven by Kim and Roush [63]; see also the survey article [125]. However, it is a complete invariant of a weaker form of conjugacy; we say that X_A and X_B are *eventually conjugate* if all sufficiently large powers are conjugate. It is not hard to show that if $A \sim B$, then X_A and X_B are eventually conjugate, and the converse is true as well:

Theorem 9 (Kim and Roush [60], Williams [131]) *Edge shifts X_A and X_B are eventually conjugate if and only if A and B are shift equivalent.*

Also the Kim–Roush counterexamples and subsequent work do not bear on a special case of the conjugacy problem: if A is shift equivalent to the 1×1 matrix $[n]$, is



Symbolic Dynamics, Figure 5
A state splitting

X_A conjugate to the full n-shift? This question, known as the *little shift equivalence problem*, is particularly intriguing because in this case the condition $A \sim [n]$ is simply the statement that A has exactly one non-zero eigenvalue, namely n.

Shift equivalence can be characterized in another way that has turned out to be very useful. Let A be an $r \times r$ integral matrix. Let R_A denote the real eventual range of A, i.e, $R_A = R^r A^r$. The *dimension group* of A is defined:

$$\Delta_A = \{v \in R_A : vA^k \in \mathbf{Z}^r \text{ for some } k \geq 0\} .$$

The *dimension group automorphism* δ_A of A is the restriction of A to Δ_A, so that $\delta_A(\mathbf{v}) = \mathbf{v}A$ for $\mathbf{v} \in \Delta_A$. The *dimension pair* of A is (Δ_A, δ_A).

If A is also nonnegative, then we define the *dimension semigroup* of A to be

$$\Delta_A^+ = \{v \in R_A : vA^k \in (Z^+)^r \text{ for some } k \geq 0\} .$$

The *dimension triple* of A is $(\Delta_A, \Delta_A^+, \delta_A)$.

It can be shown that the dimension triple completely characterizes shift equivalence, i. e., two nonnegative integral matrices are shift equivalent if and only if their dimension groups are isomorphic by an isomorphism that preserves the dimension semigroup and intertwines the dimension group automorphisms. Also, by associating equivalence classes of certain subsets of the shift space X_A to elements of Δ_A^+, one can interpret the dimension triple in terms of the action of the shift map on X_A [27,68]. And the dimension triple arises prominently in the study of the automorphism group of an SFT, which we now briefly describe. The dimension group for SFT's was developed by Krieger [68,69].

In many areas of mathematics, objects are studied by means of their symmetries. This holds true in symbolic dynamics, where symmetries are expressed by automorphisms. An *automorphism* of a shift space X is a conjugacy from X to itself. The set of all automorphisms of a shift space X is a group under composition, and is naturally called the *automorphism group*, denoted aut(X).

The goals are to understand aut(X) as a group (What kinds of subgroups does it contain? How "big" is it?) and how it acts on X, e. g., given shift-invariant subsets U, V, such as finite sets of periodic points, when is there an automorphism of X that maps U to V?. One might hope that the automorphism group would shed new light on the conjugacy problem for SFT's. Indeed, tools developed to study the automorphism group eventually paved the way for Kim and Roush to find examples of shift equivalent matrices that are not strong shift equivalent. On the other hand, the automorphism group cannot tell the entire story. For instance, aut(X_A) and aut($X_{A^\top}$) are isomorphic, since any automorphism read backwards can be viewed as an automorphism of the transposed shift, yet X_A and $X_{A^\top}$ may fail to be conjugate (for an example due to Kollmer, see p. 81 of [105]). It is not even known if the automorphism groups of the full 2-shift and the full 3-shift are isomorphic.

A good deal of our understanding of the action of the automorphism group on an SFT comes from understanding its induced representation as an action of the dimension group; this action is known as the *dimension representation*. For a much more thorough exposition on aut(X), we refer the reader to Wagoner [125,126].

Other Coding Problems

The difficulties encountered in attempts to solve the conjugacy problem motivated the formulation and study of weaker, but meaningful, notions of equivalence. For instance, we might say that two shift spaces are equivalent if one can be invertibly encoded to the other by some kind of "finite-state machine". A precise version of this is as follows.

Shift spaces X and Y are *finitely equivalent* if there is an SFT W together with finite-to-one factor codes $\phi_X : W \to X$ and $\phi_Y : W \to Y$. We call W a *common extension*, and ϕ_X, ϕ_Y the *legs*.

Here, by "finite-to-one" we mean merely that each point has a finite number of inverse images. It can be shown that any finite-to-one factor code from one shift space to another must preserve entropy, and so entropy is an invariant of finite equivalence. For irreducible sofic shifts, entropy is a complete invariant:

Theorem 10 *(Parry [100]) Two irreducible sofic shifts are finitely equivalent if and only if they have the same entropy.*

Note that from this result and the fact that finite-to-one codes between general shift spaces preserve entropy, for irreducible sofic shifts, we could have just as well defined finite equivalence with the common extension W merely being a shift space. However, if W is an SFT, we get a more concrete coding interpretation as follows. First, we recode W to an edge shift X_G and recode the legs, $\phi_X = (\Phi_X)_\infty$ and $\phi_Y = (\Phi_Y)_\infty$, to one-block codes, and (with a bit more argument) we can assume that G is irreducible. In this set-up, the finite-to-one condition translates to the so-called "no-diamond" condition, which means that for any given pair of states I, J and finite sequence w, there is at most one path from I to J with label w [33,34]. Since, for any fixed state I, the number of cycles of length n, a multiple of $p = \mathrm{per}(G)$, at I grows like $2^{nh(W)}$, we have, in this set-up a means to invertibly encode a "large" set of allowed blocks in X to allowed blocks in Y: namely, fix state I, and a large n, which is a multiple of p; for any cycle γ of length n at state I encode the Φ_X-label of γ to the Φ_Y-label of γ.

For encoding and decoding, one can dispense with state information if the legs are "almost invertible". A factor code ϕ is *almost invertible* if it is one-to-one on sequences that are typical in the following sense: x is typical if every allowed block appears infinitely often in x both to the left and the right. We then say that shift spaces X and Y are *almost conjugate* if there is an SFT W and almost invertible factor codes $\phi_X : W \to X$, $\phi_Y : W \to Y$. We call (W, ϕ_X, ϕ_Y) an *almost conjugacy* between X and Y.

For irreducible sofic shifts, it can be shown that any almost invertible factor code is finite-to-one, and so almost conjugacy implies finite equivalence. Thus, entropy is again an invariant, and together with a second very mild invariant, it is complete:

Theorem 11 *(Adler–Marcus [4]) Let X and Y be irreducible sofic shifts with minimal right resolving presentations $(G, \mathcal{L})$ and $(H, \mathcal{M})$. Then X and Y are almost conjugate if and only if $h(X) = h(Y)$ and $\mathrm{per}(G) = \mathrm{per}(H)$.*

In particular, if X and Y are mixing, then $\mathrm{per}(G) = 1 = \mathrm{per}(H)$ and entropy itself is a complete invariant.

Thus, with an an almost conjugacy, one can invertibly encode most sequences in X to those of Y without the need for auxiliary state information. Moreover, if X and Y are almost conjugate, then there is an almost conjugacy of X and Y in which one leg is right-resolving and the other leg is left-resolving (and the common extension is irreducible [4]). This gives an even more concrete interpretation to the encoding.

The proofs of Theorems 10 and 11 are actually quite constructive. For illustration we consider a very special, but historically important, case.

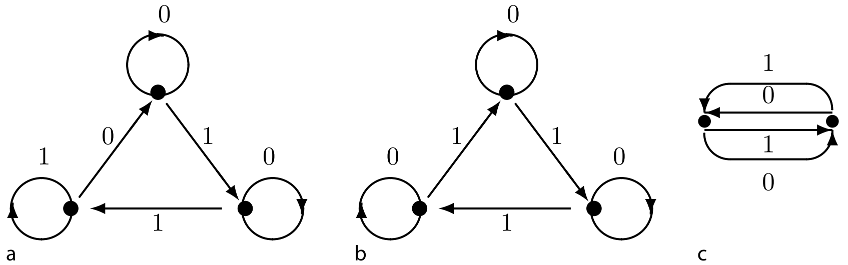
Let G be a graph with constant out-degree n. A *road coloring* Φ is a labeling of G such that at each state of G, each symbol $0, \dots, n-1$ appears exactly once as the label of an outgoing edge. An n-ary word w is *synchronizing* if all paths that are labeled w end at the same state. Figure 6 gives examples of road-colorings, with $n = 2$.

For a road-coloring, a binary word may be viewed as a sequence of instructions given to drivers starting at each of the states. A synchronizing word is a word that drives everybody to the same state. For instance, in Fig. 6a, the word 11 drives everybody to the state in the lower-left corner. But Fig. 6b does not have a synchronizing word because whenever a driver takes a '0' road he stays where he is and whenever a driver takes a '1' road he rotates by 120 degrees. The road-coloring in Fig. 6c is essentially the only road-coloring of its underlying graph, and there is no synchronizing word because drivers must always oscillate between the two states.

Now, let X be the full n-shift and Y be an irreducible SFT with entropy $\log n$. Suppose that we could find a presentation $(G, \mathcal{L})$ of Y with G having constant out-degree n and $\mathcal{L}_\infty$ finite-to-one. Then, define Φ to be any road coloring of G. Then Φ_∞ would be a finite-to-one (in fact, right resolving!) factor code from X_G to X. And we would obtain a finite equivalence with $X = X_G$, $\phi_Y = \mathcal{L}_\infty$ and $\phi_X = \Phi_\infty$. If, moreover, we could choose $\mathcal{L}$ and Φ such that ϕ_Y and ϕ_X are almost invertible, then we would have an almost conjugacy.

It turned that this could be arranged for ϕ_Y via a construction related to state splitting [2]. And if Y were mixing, then G could be chosen to be primitive and ϕ_Y almost invertible. In this setting, $\phi_X = \Phi_\infty$ would be almost invertible iff Φ has a synchronizing word; the sufficiency follows from the fact that every bi-infinite binary sequence which contains w infinitely often to the left would be the label of exactly one bi-infinite sequence of edges (to see this, use the synchronizing and road-coloring properties).

The construction of such a labeling Φ became known



Symbolic Dynamics, Figure 6
Some road-colorings

as the Road Problem, which remained open for thirty years. In the meantime, a weaker version of the the road problem was solved and applied to yield this special case of Theorem 11 (see [2]). Nevertheless, the problem remained an important problem in graph/automata theory and was only recently solved:

Theorem 12 *(Road theorem (Trachtman [122])) If G is a finite directed primitive graph with constant out-degree n, there a road-coloring of G which has a synchronizing word.*

Trachtman's approach relies heavily on earlier work of Friedman [44] and Kari [57].

The primitivity assumption above is close to necessary. Clearly some kind of connectivity is required and in the presence of irreducibility, primitivity would be necessary since otherwise there would be a "phase" introduced in the graph that would never allow a word to synchronize, as in Fig. 6c.

So far, we have focused on equivalences between symbolic systems. There has also been considerable attention paid to problems of embedding one system into another and factoring one onto another.

One of the most striking results of this type is the Krieger embedding theorem. It is not hard to show that any proper subshift of an irreducible SFT must have strictly smaller entropy. Thus, a necessary condition for a proper embedding of a shift space into an irreducible SFT is that it have strictly smaller entropy. This condition, together with a trivially necessary condition on periodic points, turns out to be sufficient. Recall that $q_n(X)$ denotes the number of points of least period n in X.

Theorem 13 *(Embedding Theorem (Krieger [70])) Let X and Y be irreducible shifts of finite type. Then there is a proper embedding of X into Y if and only if $h(X) < h(Y)$ and for each $n \geq 1$, $q_n(X) \leq q_n(Y)$.*

In fact, Krieger's theorem shows that these conditions are necessary and sufficient for a proper embedding of any shift space into a mixing shift space. The analogous problems for embedding into irreducible or mixing sofic shifts are still open, though there are partial results [22].

Using the embedding theorem and other tools from symbolic dynamics, Boyle and Handelman [25] obtained a stunning application to linear algebra: namely, a complete characterization of the non-zero spectra of primitive matrices over **R**. In fact, they obtained characterizations of non-zero spectra for primitive matrices over many other subrings of **R**. While they did not obtain a complete characterization over **Z**, they formulated a conjecture for **Z** and obtained many partial results towards that conjecture, which was later proven using other tools. The result, stated below, shows that three simple necessary conditions on a set of nonzero complex numbers are actually sufficient. In order to state these conditions, we need the following notation:

Let $\Lambda = \{\lambda_1, \ldots, \lambda_k\}$ be a list of nonzero complex numbers (with multiplicity). Let $f_\Lambda(t) = \prod_{i=1}^{k}(t - \lambda_i)$, and $\mathrm{tr}_n(\Lambda) = \sum_{d/n} \mu(n/d) \sum_{i=1}^{k} \lambda_i^k$, with μ being the Mobius Inversion function.

- Integrality Condition: $f_\Lambda(t)$ is a monic polynomial (with integer coefficients).
- Perron Condition: There is a positive entry in Λ, occurring just once, that strictly dominates in absolute value all other entries. We denote this entry by λ_Λ.
- Net Trace Condition: $\mathrm{tr}_n(\Lambda) \geq 0$ for all $n \geq 1$.

Theorem 14 *(Kim–Ormes–Roush [64]) Let Λ be a list of nonzero complex numbers satisfying the Integrality, Perron, and Net Trace Conditions. Then there is a primitive integral matrix A for which Λ is the non-zero spectrum of A.*

These conditions are all indeed necessary for Λ to be the non-zero spectrum of a primitive integral matrix. The integrality condition is that Λ forms a complete set of algebraic conjugates; the Perron condition states that Λ must satisfy the conditions of the Perron–Frobenius theorem for primitive matrices; and the Net Trace condition assures that the number of periodic points of least period n would be nonnegative.

We now turn from embeddings to factors. One special case, which is somewhat related to the Road Problem above and also important for data recording applications (Sect. "Coding for Data Recording Channels") is:

Theorem 15 [1,81] *An SFT X factors onto the full n-shift iff* $h(X) \geq \log(n)$.

While this special case treats both the equal entropy case ($h(X) = \log(n)$) and unequal entropy case ($h(X) > \log(n)$), in general, the factor problem naturally divides into two cases: lower entropy factors and equal entropy factors. In either case, a trivial necessary condition for Y to be a factor of X is that whenever $q_n(X) \neq 0$, there exists a d/n such that $q_d(Y) \neq 0$. This condition is denoted $P(X) \searrow P(Y)$. Building on ideas from Krieger's embedding theorem, this necessary condition was shown to be sufficient.

Theorem 16 *(Lower entropy factor theorem (Boyle* [22]*)) Let X and Y be irreducible SFT's with* $h(X) > h(Y)$. *Then there is a factor code from X to Y if and only if* $P(X) \searrow P(Y)$.

As with the embedding theorem, the lower entropy factors problem for irreducible sofic shifts is still open.

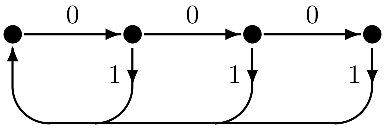
The equal entropy factors problem for SFT's is quite different. Clearly, $P(X) \searrow P(Y)$ is a necessary condition. A second necessary condition involves the dimension group and is simplest to state in the case of mixing edge shifts X_A and X_B.

We say that a subgroup Δ of the dimension group Δ_A is *pure* if whenever an integer multiple of an element $v \in \Delta_A$ is in Δ, then so is v; intuitively Δ does not have any "rational holes" in Δ_A. The condition is that there is a pure δ_A-invariant subgroup Δ of Δ_A such that (Δ_B, δ_B) is a quotient of $(\Delta, \delta_A|_\Delta)$.

In the equal entropy case, this condition and the trivial periodic point condition, $P(X) \searrow P(Y)$, subsume all known necessary conditions for the existence of a factor code from one mixing edge shift to another. It is not known if these two conditions are sufficient. For references on this problem, see [9,27,66].

Coding for Data Recording Channels

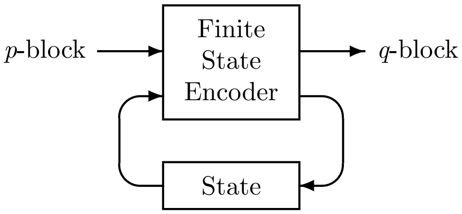
In magnetic recording, within any given clock cell, a '1' is represented as a change in magnetic polarity, while a '0' is represented as an absence of such a change. Two successive 1's (separated by some number, $m \geq 0$, of 0's) are read as a voltage peak followed by a voltage trough (or vice versa). If the peak and trough occur too close together, *intersymbol interference* can occur: the amplitudes of the peak and trough are degraded, and the positions at which they occur are distorted. In order to control intersymbol interference, it is desirable that 1's not be too close together, or equivalently that runs of 0's not be too short. On the other hand, for timing control, it is desirable that runs of 0's not be too long; this is a consequence of the fact that only 1's are observed: the length of a run of 0's is inferred by connection to a clock via a feedback loop, and a long run of 0's could cause the clock to drift more than one clock cell.



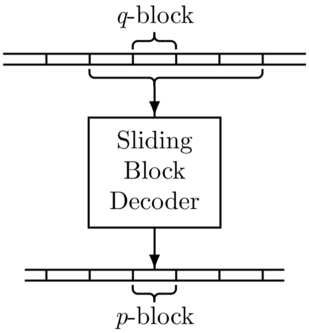
Symbolic Dynamics, Figure 7
$X(1,3)$

This gives rise to *run-length-limited shift spaces*, $X(d,k)$, where runs of 0's are constrained to be bounded below by some positive integer d and bounded above by some positive $k \geq d$. More precisely, $X(d,k)$ is defined by the constraints that 1's occur infinitely often in each direction, and there are at least d 0's, but no more than k 0's, between successive 1's. Note that $X(d,k)$ is an SFT with forbidden list $\mathcal{F} = \{0^{k+1}, 10^i1, 0 \leq i < d\}$. Figure 7 depicts a labeled graph presentation of $X(1,3)$. The SFT's $X(1,3)$, $X(2,7)$, and $X(2,10)$ have been used in floppy disks, hard disks, and the compact audio disk, respectively.

Now in order to record completely arbitrary information, we need to build an encoder which encodes arbitrary binary sequences into sequences that satisfy a given constraint (such as $X(d,k)$). The encoder is a finite-state machine, as depicted in Fig. 8. It maps arbitrary binary data sequences, grouped into blocks of length p (called p-blocks), into constrained sequences, grouped into blocks of length q (called q-blocks). The encoded q-block is a function of the p-block as well as an internal state. When concatenated together, the sequence of encoded q-blocks must satisfy the given constraint. Also, the encoded sequences should be decodable, meaning that given the initial encoder state, a string of p-blocks can be



Symbolic Dynamics, Figure 8
Encoder



Symbolic Dynamics, Figure 9
Sliding block decoder

recovered from its encoded string of q-blocks, possibly allowing a fixed delay in time.

If the constrained sequences satisfy the constraints of a sofic shift X, we say that such a code is a *rate p:q finite-state code into X*. In terms of symbolic dynamics, such a code consists of an edge shift X_G, a right resolving factor code ϕ_1 from X_G onto the full 2^p-shift and a right closing sliding block code ϕ_2 into X^q (the right closing condition expresses the decodability condition).

In most applications, it is important that a stronger decoding condition be imposed. Namely, a *sliding block decodable* rate p:q finite-state code into X consists of a finite-state code given by (X_G, ϕ_1, ϕ_2) and a sliding block code $\psi: X^q \to X_{2^p}$ such that $\psi \circ \phi_2 = \phi_1$. This means that the decoded p-block depends only upon the local context of the received q-block – that is, decoding is accomplished by applying a time-invariant function to a window consisting of a bounded amount of memory and/or anticipation, but otherwise is state-independent (see Fig. 9, which depicts the situation where the memory is 1 and the anticipation is 2). The point is that whenever the window of the decoder passes beyond a raw channel error, that error cannot possibly affect future decoding; thus, sliding block decoders control error propagation.

Symbolic dynamics has played an important role in providing a framework for constructing such codes as well as for establishing bounds on various figures of merit for such codes. Specifically, by modifying the constructions used in the proofs of Theorems 10 and 11, in the early 1980's, Adler, Coppersmith and Hassner (ACH) established the following results:

Theorem 17 *(Finite-state coding theorem* [1]*) Let X be a sofic shift and p, q be positive integers. Then there is a rate p:q finite-state code into X if and only if $p/q \le h(X)$.*

Theorem 18 *(Sliding block decoding theorem* [1]*) Let X be an SFT and p, q be positive integers. Then there is a rate p:q sliding-block decodable finite-state code into X if and only if $p/q \le h(X)$.*

We have described all of this in the context of the binary alphabet for data sequences. In fact, it works just as well for any finite alphabet, and the method used to prove Theorem 18 solved, at the same time, a special case of the factor problem for SFT's: namely, Theorem 15 above. Sometimes, Theorems 17 and 18 are stated only for rate 1 : 1 codes but in the context of arbitrary finite alphabets (e. g. see Chap. 5 in [78]), this easily extends to the general $p : q$ case by passing to powers.

The ACH paper was the beginning of a rigorous theory of constrained coding. Theorem 18 has been extended to a large class of sofic shifts, and bounds have been established on such figures of merit as number of encoder states as well as the size of the decoding window of the sliding block decoder. While the state-splitting algorithm does construct codes with "relatively small" decoding windows, it is not yet understood how to construct codes with the smallest such windows. This is of substantial engineering interest since the smaller the decoding window, the smaller the error propagation. There is now a substantial literature on the construction of these types of codes, including the state-splitting algorithm as well as many other methods of encoder/decoder design and a wealth of examples that go well beyond the run length limited constraints. See for example the expositions [11,54], and [83] as well as the papers [7,8,10,12,31,41,42,43,53,56].

Connections with Information Theory and Ergodic Theory

The concept of entropy was developed by Shannon [119] in information theory in the 1940's and was adapted to ergodic theory in the 1950's and to dynamical systems, and in particular symbolic dynamics, in the 1960's. Shannon focused on entropy for random variables and finite sequences of random variables, but the concept naturally extends to stationary stochastic processes, in particular stationary Markov chains. Roughly speaking, for a stationary process μ, the entropy $h(\mu)$ is the asymptotic growth rate of the number of allowed sequences, weighted by the joint stationary probabilities of μ (for background on entropy and information theory see Cover and Thomas [35]).

A Markov chain on a graph G is defined by assigning transition probabilities to the edges. Let P denote the stochastic matrix indexed by states of G, with P_{IJ} equal to the sum of the transition probabilities of all edges from I to J. This, together with a stationary vector for P, completely defines the (joint distributions of the) Markov chain.

So, a graph itself can be viewed as specifying only which transitions are possible. For that reason, one could view G as an "intrinsic Markov chain", and Parry [99] used this terminology when he introduced SFT's based on graphs and matrices.

More generally, a stationary process on a shift space X is any stationary process which assigns positive probability only to allowed blocks in X.

For an irreducible graph G with strictly positive transition probabilities on all edges, by Perron–Frobenius theory, P will always have a unique stationary vector. And there is a particular such Markov chain μ_G on G that in some sense distributes probabilities on paths as uniformly as possible. This Markov chain is the most "random" possible stationary process (not just among Markov chains) on X_G in the sense that it has maximal entropy; moreover, its entropy coincides with the topological entropy of X_G.

The Markov chain μ_G is defined as follows: let λ denote the largest eigenvalue of $A(G)$ and w, v denote corresponding left, right eigenvectors, normalized such that $w \cdot v = 1$; for any path γ of length n from state I to state J, we define the stationary probability of γ:

$$\mu_G(\gamma) = \frac{w_I v_J}{\lambda^n} .$$

It is clear from the formula that this distribution is fairly uniform, since all paths with the same initial state, terminal state and length have the same stationary probability.

This defines a joint stationary distribution on allowed blocks of X_G, and it is not hard to show that it is consistent and Markov, given by assigning transition probability $v_J/(v_I \lambda)$ to each edge from I to J. To summarize:

Theorem 19 *Let G be an irreducible graph.*

- *Any stationary process* μ *on G satisfies* $h(\mu) \le \log \lambda$, *and*
- *The Markov chain* μ_G *is the unique stationary process on G such that* $h(\mu_G) = \log \lambda$.

The construction of μ_G is effectively due to Shannon [119] and uniqueness is due to Parry [99] The unique entropy-maximizing stationary process on irreducible edge shifts naturally extends to irreducible SFT's and irreducible sofic shifts (this process will be M-step Markov for an M-step SFT).

Many results in symbolic dynamics were originally proved using μ_G. One example is the fact that any factor code from one irreducible SFT to another of the same entropy must be finite-to-one [32]. This result, and many others, were later proven using methods that rely only on the basic combinatorial structure of G and X_G. Nevertheless, μ_G provides much motivation and insight into symbolic dynamics problems, and it illustrates a connection with ergodic theory, which we now discuss.

For background on ergodic theory (such as the concepts of measure-preserving transformations, homomorphisms, isomorphisms, ergodicity, mixing, and measure-theoretic entropy), we refer the reader to [107,113,127]. Observe that a stationary process on a shift space X can be viewed as a measure-preserving transformation: the transformation is the shift mapping and the measure on "cylinder sets", consisting of $x \in X$ with prescribed coordinate values $x_i = a_i, \ldots, x_j = a_j$, is defined as the probability of the word $a_i \ldots a_j$; the stationarity of the process translates directly into preservation of the measure. The symbolic dynamical concepts of irreducibility and mixing correspond naturally to the concepts of ergodicity and (measure-theoretic) mixing in ergodic theory.

It is well-known [107,127] that the measure-preserving transformation (MPT) defined by a stationary Markov chain μ on an irreducible (resp., primitive) graph G is ergodic (resp., mixing) if μ assigns strictly positive conditional probabilities to all edges of G.

Now, suppose that G and H are irreducible graphs and $\phi\colon X_G \to X_H$ is a factor code. For a stationary measure μ on G, we define a stationary measure $\nu = \phi(\mu)$ on X_H by transporting μ to X_H: for a measurable set A in X_H, define

$$\nu(A) = \mu\left(\phi^{-1}(A)\right) .$$

Then ϕ defines a measure-preserving homomorphism from the MPT defined by μ to the MPT defined by ν. Since measure-preserving homomorphisms between MPT's cannot reduce measure-theoretic entropy, we have $h(\nu) \le h(\mu)$.

Suppose that now ϕ is actually a conjugacy. Then it defines a measure-preserving isomorphism, and so $h(\mu) = h(\nu)$. If $\mu = \mu_G$, then by uniqueness, we have $\nu = \mu_H$. Thus ϕ defines a measure-theoretic isomorphism between the MPT defined by μ_G and the MPT defined by μ_H. In fact, this holds whenever ϕ is merely an almost invertible factor code. This establishes the following result:

Theorem 20 *Let* G, H *be irreducible graphs, and let* μ_G, μ_H *be the stationary Markov chains of maximal entropy on* G, H. *If* X_G, X_H *are almost conjugate (in particular, if they are conjugate), then the measure-preserving transformations defined by* μ_G *and* μ_H *are isomorphic.*

Hence conjugacies and almost conjugacies yield isomorphisms between measure-preserving transformations defined by stationary Markov chains of maximal entropy.

In fact, the isomorphisms obtained in this way have some very desirable properties compared to the run-of-the-mill isomorphism. For instance, an isomorphism ϕ between stationary processes typically has an infinite window; i. e., to know $\phi(x)_0$, you typically need to know all of x, not just a central block $x_{[-n,n]}$ (these are the kinds of isomorphisms that appear in general ergodic theory and in particular in Ornstein's celebrated isomorphism theory [98]). In contrast, by definition, a conjugacy always has a finite window of uniform size. It turns out that an isomorphism obtained from an almost conjugacy, as well as its inverse, has finite expected coding length in the sense that to know $\phi(x)_0$, you need to know only a central block $x_{[-n(x),n(x)]}$, where the function $n(x)$ has finite expectation [4]. In particular, by Theorem 11, whenever X_G and X_H are mixing edge shifts with the same entropy, the measure-preserving transformations defined by μ_G and μ_H are isomorphic via an isomorphism with finite expected coding length.

The notions of conjugacy, finite equivalence, almost conjugacy, embedding, factor code, and so on can all be generalized to the context of stationary measures, in particular to stationary Markov chains. For instance, a conjugacy between two stationary measures is a map that is simultaneously a conjugacy of the underlying shift spaces and an isomorphism of the associated measure-preserving transformations. Many results in symbolic dynamics have been generalized to the context of stationary Markov chains. There is a substantial literature on this, in particular on finitary isomorphisms with finite expected coding time, e. g., [71,84,101,114], and [90]. The expositions [102,105] give a nice introduction to the subject of strong finitary codings between stationary Markov chains. See also the research papers [45,85,86,103,104,123,124].

Higher Dimensional Shift Spaces

In this section we introduce higher dimensional shift spaces. For a more thorough introduction, we refer the reader to Lind [76]. For the related subject of tiling systems, see Robinson [112], Radin [110], and Mozes [94,95].

The *d-dimensional full $\mathcal{A}$-shift* is defined to be $\mathcal{A}^{\mathbf{Z}^d}$. Ordinarily, $\mathcal{A}$ is a finite alphabet, and here we restrict ourselves to this case. An element x of the full shift may be regarded as a function $x\colon \mathbf{Z}^d \to \mathcal{A}$, or, more informally, as a "configuration" of alphabet choices at the sites of the integer lattice $\mathbf{Z}^d$.

For $x \in \mathcal{A}^{\mathbf{Z}^d}$ and $F \subseteq \mathbf{Z}^d$, let x_F denote the restriction of x to F. The usual metric on the one-dimensional full shift naturally generalizes to a metric on $\mathcal{A}^{\mathbf{Z}^d}$ given by $\rho(x, y) = 2^{-k}$, where k is the largest integer such that $x_{[-k,k]^d} = y_{[-k,k]^d}$ (with the usual conventions when $x = y$ and $x_{\mathbf{0}} \neq y_{\mathbf{0}}$). In analogy with one dimension, according to this definition, two points are "close" if they agree on a large cube $[-k, k]^d$.

We define higher dimensional shift spaces, with the following terminology. A *shape* is a finite subset F of $\mathbf{Z}^d$, and a *pattern f* on a shape F is a function $f\colon F \to \mathcal{A}$. We say that X is a *d-dimensional shift space* (or *d-dimensional shift*) if it can be represented by a list $\mathcal{F}$ (finite or infinite) of "forbidden" patterns

$$X = X_{\mathcal{F}} = \big\{x \in \mathcal{A}^{\mathbf{Z}^d} : \sigma^n(x)_F \notin \mathcal{F} \\ \text{for all } n \in \mathbf{Z}^d \text{ and all shapes } F\big\} .$$

Just as in one dimension, we can equivalently define a shift space to be a closed (with respect to the metric ρ) translation-invariant subset of $\mathcal{A}^{\mathbf{Z}^d}$. Here "translation-invariance" means that $\sigma^n(X) = X$ for all $n \in \mathbf{Z}^d$, where σ^n is the translation in direction n defined by $(\sigma^n(x))_m = x_{m+n}$.

We say that a pattern f on a shape F *occurs* in a shift space X if there is an $x \in X$ such that $x_F = f$. Hence the analogue of the language of a shift space is the set of all occurring patterns.

A *d-dimensional shift of finite type* X is a subset of $\mathcal{A}^{\mathbf{Z}^d}$ defined by a finite list $\mathcal{F}$ of forbidden patterns. Just as in one dimension, a d-dimensional shift of finite type X can also be defined by specifying allowed patterns instead of forbidden patterns, and there is no loss of generality in requiring the shapes of the patterns to be the same. Thus we can specify a finite list $\mathcal{L}$ of patterns on a fixed shape F, and set

$$X = X_{\mathcal{L}^c} = \big\{x \in \mathcal{A}^{\mathbf{Z}^d} : \text{ for all } n \in \mathbf{Z}^d\,,\ \sigma^n(x)_F \in \mathcal{L}\big\} .$$

In fact, there is no loss in generality in assuming that F is a d-dimensional cube $F = [0, k]^d$.

Given a finite list $\mathcal{L}$ of patterns on a shape F, we say that a pattern f' on a shape F' is *$\mathcal{L}$-admissible* (in the shift of finite type $X = X_{\mathcal{L}^c}$) if each of its sub-patterns, whose shape is a translate of F, belongs to $\mathcal{L}$. Of course, any pattern which occurs in X is $\mathcal{L}$-admissible. But an $\mathcal{L}$-admissible pattern need not occur in X.

The analogue of vertex shift (or 1-step shift of finite type) in higher dimensions is defined by a collection of d transition matrices $A_1, \ldots, A_d$ all indexed by the same set of symbols $\mathcal{A} = \{1, \ldots, m\}$. We set

$$\Omega(A_1, \ldots, A_d) \\ = \big\{x \in \{1, \ldots, m\}^{\mathbf{Z}^d} : A_i(x_n, x_{n+e_i}) = 1 \text{ for all } n, i\big\},$$

where e_i is as usual the ith standard basis vector and $A_i(a, b)$ denotes the (a, b)-entry of A_i. Such a shift space is called a *matrix subshift*. When $d = 2$, this amounts to a pair of transition matrices A_1 and A_2 with identical vertex sets. The matrix A_1 controls transitions in the horizontal direction and the matrix A_2 controls transitions in the vertical direction. Note that any matrix subshift is a shift of finite type and, in particular, can be specified by a list $\mathcal{L}$ of patterns on the unit cube $F = \{(a_1, \dots, a_n)\colon a_i \in \{0, 1\}\}$; specifically, $\Omega(A_1, \dots, A_n) = X_{\mathcal{L}^c}$ where $\mathcal{L}$ is the set of all patterns $f\colon F \to \{1, \dots, m\}$ such that if $n, n + e_i \in F$, then $A_i(f(n), f(n + e_i)) = 1$. When we speak of admissible patterns for a matrix subshift, we mean $\mathcal{L}$-admissible patterns with this particular $\mathcal{L}$. Just as in one dimension, we can recode any shift of finite type to a matrix subshift.

Higher dimensional SFT's can behave very differently from one dimension. For example, there is a simple method to determine if a one-dimensional edge shift, and therefore a one-dimensional shift of finite type, is nonempty, and there is an algorithm to tell, for a given finite list $\mathcal{L}$, whether a given block occurs in $X = X_{\mathcal{L}^c}$. The corresponding problems in higher dimensions, called the *nonemptiness problem* and the *extension problem*, turn out to be undecidable [14,111]; see also [67]. Even for two-dimensional matrix subshifts X, these decision problems are undecidable. On the other hand, there are some special classes where these problems are decidable. This class includes any two-dimensional matrix subshift such that A_1 commutes with A_2 and $A_2^{\top}$. For this class, any admissible pattern on a cube must occur, and so the nonemptiness and extension problems are decidable; see [87,88].

A point x in a d-dimensional shift X is *periodic* if its orbit $\{\sigma^n(x)\colon n \in \mathbf{Z}^d\}$ is finite. Observe that this reduces to the usual notion of periodic point in one dimension. Now an ordinary (one-dimensional) nonempty shift of finite type is conjugate to an edge shift X_G, where G has at least one cycle. Hence a one-dimensional shift of finite type is nonempty if and only if it has a periodic point. This turns out to be false in higher dimensions [14,111], and this fact is intimately related to the undecidability results mentioned above. While one can formulate a notion of zeta function for keeping track of numbers of periodic points, the zeta function is hard to compute, even for very special and explicit matrix subshifts, and, even in this setting, it is not a rational function[77].

In higher dimensions, the entropy of a shift is defined as the asymptotic growth rate of the number of occurring patterns in arbitrarily large cubes. In particular, for two-dimensional shifts it is defined by

$$h(X) = \lim_{n\to\infty} \frac{1}{n^2} \log |X_{[0,n-1]\times[0,n-1]}| ,$$

where $X_{[0,n-1]\times[0,n-1]}$ denotes the set of occurring patterns on the square

$$[0, n-1] \times [0, n-1]$$

that occur in X. Recall from Sect. "Entropy and Periodic Points" that it is easy to compute the entropy of a (one-dimensional) shift of finite type using linear algebra. But in higher dimensions, there is no analogous formula and, in fact, other than the group shifts mentioned below, the entropies of only a very few higher dimensional shifts of finite type have been computed explicitly. Even for the two-dimensional "golden mean" matrix subshift defined by the horizontal and vertical transition matrices

$$A_1 = A_2 = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$$

an explicit formula for the entropy is not known. However, there are good numerical approximations to the entropy of some matrix subshifts (e. g., [30,96]). And recently the set of numbers that can occur as entropies of SFT's in higher dimensions has been characterized [51,52]; this characterization turns out to be remarkably different from the analogous characterization in one dimension [74].

One of the few 2-dimensional SFT's for which entropy has been computed is the domino tiling system (see [58], Chap. 5 in [115]), which consists of all possible tilings of the plane using the 1 × 2 and 2 × 1 dominoes. This can be translated into an SFT with four symbols L, R, T, B, subject to the constraints:

- $x_{i,j} = L \Rightarrow x_{i+1,j} = R$,
- $x_{i,j} = R \Rightarrow x_{i-1,j} = L$,
- $x_{i,j} = T \Rightarrow x_{i,j-1} = B$,
- $x_{i,j} = B \Rightarrow x_{i,j+1} = T$.

The entropy of this SFT is given by a remarkable integral formula:

$$\frac{1}{4}\int_0^1 \int_0^1 (4 - 2\cos(2\pi s) - 2\cos(2\pi t))\, \mathrm{d}s\, \mathrm{d}t$$

Further work along these lines can be found in [59,117].

One can formulate notions of irreducibility and mixing for higher dimensional shift spaces. It turns out that for SFT's there are several notions of mixing that all coincide in one dimension, but are vastly different in higher dimensions. For instance, a higher dimensional shift space is *strongly irreducible* if there is an integer $R > 0$ such that for any two shapes F, F' of distance at least R, any occurring configurations on F and F' can be combined to form an occuring configuration on $F \cup F'$ [29,129]. For one-dimensional SFT's, this is equivalent to mixing, but much stronger than mixing for two-dimensional SFT's.

Just as in one dimension, we have the notion of sliding block code for higher dimensional shifts. For finite alphabets $\mathcal{A}, \mathcal{B}$, a cube $F \subset \mathbf{Z}^d$ and a function $\Phi\colon \mathcal{A}^F \to \mathcal{B}$, the mapping $\phi = \Phi_\infty\colon \mathcal{A}^{Z^d} \to \mathcal{B}^{Z^d}$ defined by

$$\Phi_\infty(x)_n = \Phi(x_{n+F})$$

is called a sliding block code. By restriction we have the notion of sliding block code from one d-dimensional shift space to another. As expected, for d-dimensional shifts X and Y, the sliding block codes $\phi\colon X \to Y$ coincide exactly with the continuous translation-commuting maps from X to Y, i. e., the maps which are continuous with respect to the metric ρ, defined above, and which satisfy $\phi \circ \sigma^n = \sigma^n \circ \phi$ for all $n \in \mathbf{Z}^d$. Thus it makes sense to consider the various coding problems, in particular the conjugacy, factor and embedding problems, in the higher dimensional setting, but these are very difficult.

Even the question of determining when a higher dimensional SFT of entropy at least $\log n$ factors onto the full n-shift seems very difficult (in contrast to Theorem 15). However, there are some positive results for strongly irreducible SFT's. For instance, it is known that any strongly irreducible SFT of entropy strictly larger than $\log n$ factors onto the full n-shift [37,55]. In fact, that result requires only a weaker assumption than strong irreducibility, but still much stronger than mixing; recent examples [28] show, among other things, that one cannot weaken that assumption to mere mixing.

There are results on other coding problems as well. For instance, a version of the Embedding theorem in one dimension (Theorem 13) holds for SFT's in two dimensions with a strong mixing property [73]; however, it is required that the shift to be embedded contains no points that are periodic in any single direction. And some other results on entropy of proper subshifts of strongly irreducible SFT's carry over from one dimension to higher dimensions [106,109]. But in many cases where versions of the result carry over, the proofs are much different from those in one dimension.

Measures of maximal entropy for two-dimensional SFT's behave very differently from the one dimensional case. For instance, even with very strong mixing properties, such as strong irreducibility, there can be more than one measure of maximal entropy [29], and the relationships among entropy-preserving, finite-to-one, and almost invertibility for factor codes discussed in Sect. "Other Coding Problems" can be very different in higher dimensions [89]. Other differences with respect to entropy can be found in [108].

There is also the natural notion of higher dimensional sofic shifts, which can be defined as those shift spaces that are factors of SFT's. Recall that every one-dimensional sofic shift is a right-resolving, and hence entropy-preserving, factor of an SFT. It is not known if there is an analogue to this fact in higher dimensions, although recently there has been some progress: every sofic shift Y is a factor of an SFT with entropy arbitrarily close to $h(Y)$ [36].

Finally, there is a subclass of d-dimensional SFT's which is somewhat tractable, namely d-dimensional shifts with group structure in the following sense. Let $\mathcal{A}$ be a (finite) group. Then the full d-dimensional shift over $\mathcal{A}$ is also a group with respect to the coordinate-wise group structure. A (higher-dimensional) *group shift* is a subshift of $\mathcal{A}^{\mathbf{Z}^d}$ which is also a subgroup. For a survey on results for this class, we refer the reader to [79].

Future Directions

The future directions of the subject will likely be determined by progress on solutions to open problems. In the course of describing topics in this article, we have mentioned many open problems along the way. For a much more complete list on a wealth of sub-areas of symbolic dynamics, we refer the reader to the article [24]. While it is difficult to single out the most important challenges, certainly the problem of understanding multi-dimensional shift spaces, especially of finite type, is one of the most important.

Bibliography

1. Adler RL, Coppersmith D, Hassner M (1983) Algorithms for sliding block codes – an application, of symbolic dynamics to information theory. Trans IEEE Inf Theory 29:5–22
2. Adler RL, Goodwyn LW, Weiss B (1977) Equivalence of topological Markov shifts. Isr J Math 27:48–63
3. Adler R, Konheim A, McAndrew M (1965) Topological entropy. Trans Amer Math Soc 114:309–319
4. Adler R, Marcus B (1979) Topological entropy and equivalence of dynamical systems. Mem Amer Math Soc 219. AMS, Providence
5. Adler R, Weiss B (1970) Similarity of automorphisms of the torus. Mem Amer Math Soc 98. AMS, Providence
6. Aho AV, Hopcroft JE, Ullman JD (1974) The design and analysis of computer algorithms. Addison-Wesley, Reading
7. Ashley J (1988) A linear bound for sliding block decoder window size. Trans IEEE Inf Theory 34:389–399
8. Ashley J (1996) A linear bound for sliding block decoder window size, II. Trans IEEE Inf Theory 42:1913–1924
9. Ashley J (1991) Resolving factor maps for shifts of finite type with equal entropy. Ergod Theory Dynam Syst 11:219–240
10. Béal M-P (1990) The method of poles: a coding method for constrained channels. Trans IEEE Inf Theory 36:763–772
11. Béal M-P (1993) Codage symbolique. Masson, Paris
12. Béal M-P (2003) Extensions of the method of poles for code construction. Trans IEEE Inf Theory 49:1516–1523
13. Bedford T (1986) Generating special Markov partitions for hyperbolic toral automorphisms using fractals. Ergod Theory Dynam Syst 6:325–333

14. Berger R (1966) The undecidability of the Domino problem. Mem Amer Math Soc. AMS, Providence
15. Berstel J, Perrin D (1985) Theory of codes. Academic Press, New York
16. Blanchard P, Devaney R, Keen L (2004) Complex dynamics and symbolic dynamics. In: Williams S (ed) Symbolic dynamics and its applications. Proc Symp Appl Math. AMS, Providence, pp 37–59
17. Blanchard F, Maass A, Nogueira A (2000) Topics in symbolic dynamics and applications. In: Blanchard F, Maass A, Nogueira A (eds) LMS Lecture Notes, vol 279. Cambridge University Press, Cambridge
18. Bowen R (1970) Markov partitions for Axiom A diffeomorphisms. Amer J Math 92:725–747
19. Bowen R (1973) Symbolic dynamics for hyperbolic flows. Amer J Math 95:429–460
20. Bowen R, Franks J (1977) Homology for zero-dimensional basic sets. Ann Math 106:73–92
21. Bowen R, Lanford OE (1970) Zeta functions of restrictions of the shift transformation. Proc Symp Pure Math AMS 14:43–50
22. Boyle M (1983) Lower entropy factors of sofic systems. Ergod Theory Dynam Syst 3:541–557
23. Boyle M (1993) Symbolic dynamics and matrices. In: Brualdi R et al (eds) Combinatorial and graph theoretic problems in linear algebra. IMA Vol Math Appl 50:1–38
24. Boyle M (2007) Open problems in symbolic dynamics. Contemporary Math (to appear)
25. Boyle M, Handelman D (1991) The spectra of nonnegative matrices via symbolic dynamics. Ann Math 133:249–316
26. Boyle M, Krieger W (1986) Almost Markov and shift equivalent sofic systems. In: Aleixander J (ed) Proceedings of Maryland Special Year in Dynamics 1986–87. Lecture Notes in Math, vol 1342. Springer, Berlin, pp 33–93
27. Boyle M, Marcus BH, Trow P (1987) Resolving maps and the dimension group for shifts of finite type. Mem Amer Math Soc 377. AMS, Providence
28. Boyle M, Pavlov R, Schraudner M (2008) preprint
29. Burton R, Steif J (1994) Nonuniqueness of measures of maximal entropy for subshifts of finite type. Ergod Theory Dynam Syst 14(2):213–235
30. Calkin N, Wilf H (1998) The number of independent sets in a grid graph. SIAM J Discret Math 11:54–60
31. Cidecyian R, Evangelos E, Marcus B, Modha M (2001) Maximum transition run codes for generalized partial response channels. IEEE J Sel Area Commun 19:619–634
32. Coven E, Paul M (1974) Endomorphisms of irreducible shifts of finite type. Math Syst Theory 8:167–175
33. Coven E, Paul M (1975) Sofic systems. Isr J Math 20:165–177
34. Coven E, Paul M (1977) Finite procedures for sofic systems. Monats Math 83:265–278
35. Cover T, Thomas J (1991) Elements of information theory. Wiley, New York
36. Desai A (2006) Subsystem entropy for Z^d sofic shifts. Indag Math 17:353–360
37. Desai A (2008) A class of Z^d-subshifts which factor onto lower entropy full shifts. Proc Amer Math Soc, to appear
38. Devaney R (1987) An introduction to chaotic dynamical systems. Addison-Wesley, Reading
39. Fischer R (1975) Sofic systems and graphs. Monats Math 80:179–186
40. Fischer R (1975) Graphs and symbolic dynamics. Colloq Math Soc János Bólyai: Top Inf Theory 16:229–243
41. Franaszek PA (1968) Sequence-state coding for digital transmission. Bell Syst Tech J 47:143–155
42. Franaszek PA (1982) Construction of bounded delay codes for discrete noiseless channels. J IBM Res Dev 26:506–514
43. Franaszek PA (1989) Coding for constrained channels: a comparison of two approaches. J IBM Res Dev 33:602–607
44. Friedman J (1990) On the road coloring problem. Proc Amer Math Soc 110:1133–1135
45. Gomez R (2003) Positive K-theory for finitary isomoprhisms of Markov chains. Ergod Theory Dynam Syst 23:1485–1504
46. Hadamard J (1898) Les surfaces a courbures opposées et leurs lignes geodesiques. J Math Pure Appl 4:27–73
47. Hassellblatt B, Katok A (1995) Introduction to the modern theory of dynamical systems. Cambridge University Press, Cambridge
48. Hedlund GA (1939) The dynamics of geodesic flows. Bull Amer Math Soc 45:241–260
49. Hedlund GA (1944) Sturmian minimal sets. Amer J Math 66:605–620
50. Hedlund GA (1969) Endomorphisms and automorphisms of the shift dynamical system. Math Syst Theory 3:320–375
51. Hochman M, Meyerovitch T (2007) A characterization of the entropies of multidimensional shifts of finite type. Ann Math, to appear
52. Hochman M (2007) On the dynamics and recursive properties of multidimensional symbolic systems. Preprint
53. Hollmann HDL (1995) On the construction of bounded-delay encodable codes for constrained systems. Trans IEEE Inf Theory 41:1354–1378
54. Immink KAS (2004) Codes for mass data storage, 2nd edn. Shannon Foundation Press, Eindhoven
55. Johnson A, Madden K (2005) Factoring higher-dimensional shifts of finite type onto full shifts. Ergod Theory Dynam Syst 25:811–822
56. Karabed R, Siegel P, Soljanin E (1999) Constrained coding for binary channels with high intersymbol intereference. Trans IEEE Inf Theory 45:1777–1797
57. Kari J (2001) Synchronizing finite automata on Eulerian digraphs. Springer Lect Notes Comput Sci 2136:432–438
58. Kastelyn PW (1961) The statistics of dimers on a lattice. Physica A 27:1209–1225
59. Kenyon R (2008) Lectures on dimers. http://www.math.brown.edu/~rkenyon/papers/dimerlecturenotes.pdf
60. Kim KH, Roush FW (1979) Some results on decidability of shift equivalence. J Comb Inf Syst Sci 4:123–146
61. Kim KH, Roush FW (1988) Decidability of shift equivalence. In: Alexander J (ed) Proceedings of Maryland special year in dynamics 1986–87. Lecture Notes in Math, vol 1342. Springer, Berlin, pp 374–424
62. Kim KH, Roush FW (1990) An algorithm for sofic shift equivalence. Ergod Theory Dynam Syst 10:381–393
63. Kim KH, Roush FW (1999) Williams conjecture is false for irreducible subshifts. Ann Math 149:545–558
64. Kim KH, Ormes N, Roush F (2000) The spectra of nonnegative integer matrices via formal power series. Amer J Math Soc 13:773–806
65. Kitchens B (1998) Symbolic dynamics: One-sided, two-sided and countable state Markov chains. Springer, Berlin
66. Kitchens B, Marcus B, Trow P (1991) Eventual factor maps

and compositions of closing maps. Ergod Theory Dynam Syst 11:85–113

67. Kitchens B, Schmidt K (1988) Periodic points, decidability and Markov subgroups, dynamical systems. In: Alexander JC (ed) Proceedings of the special year. Springer Lect Notes Math 1342:440–454
68. Krieger W (1980) On a dimension for a class of homeomorphism groups. Math Ann 252:87–95
69. Krieger W (1980) On dimension functions and topological Markov chains. Invent Math 56:239–250
70. Krieger W (1982) On the subsystems of topological Markov chains. Ergod Theory Dynam Syst 2:195–202
71. Krieger W (1983) On the finitary isomorphisms of Markov shifts that have finite expected coding time. Wahrscheinlichkeitstheorie Z 65:323–328
72. Krieger W (1984) On sofic systems I. Isr J Math 48:305–330
73. Lightwood S (2003/04) Morphisms form non-periodic Z^2 subshifts I and II. Ergod Theory Dynam Syst 23:587–609, 24: 1227–1260
74. Lind D (1984) The entropies of topological Markov shifts and a related class of algebraic integers. Ergod Theory Dynam Syst 4:283–300
75. Lind D (1989) Perturbations of shifts of finite type. SIAM J Discret Math 2:350–365
76. Lind D (2004) Multi-dimensional symbolic dynamics. In: Williams S (ed) Symbolic dynamics and its applications. Proc Symp Appl Math 60:81–120
77. Lind D (1996) A zeta function for Z^d-actions. In: Pollicott M, Schmidt K (eds) Proceedings of Warwick Symposium on Z^d-actions. LMS Lecture Notes, vol 228. Cambridge University Press, Cambridge, pp 433–450
78. Lind D, Marcus B (1995) An introduction to symbolic dynamics and coding. Cambridge University Press, Cambridge
79. Lind D, Schmidt K (2002) Symbolic and algebraic dynamical systems. In: Hasselblatt B, Katok A (eds) Handbook of Dynamics Systems. Elsevier, Amsterdam, pp 765–812
80. Manning A (1971) Axiom A diffeomorphisms have rational zeta functions. Bull Lond Math Soc 3:215–220
81. Marcus B (1979) Factors and extensions of full shifts. Monats Math 88:239–247
82. Marcus BH, Roth RM (1991) Bounds on the number of states in encoder graphs for input-constrained channels. Trans IEEE Inf Theory 37:742–758
83. Marcus BH, Roth RM, Siegel PH (1998) Constrained systems and coding for recording chapter. In: Brualdi R, Huffman C, Pless V (eds) Handbook on coding theory. Elsevier, New York; updated version at http://www.math.ubc.ca/~marcus/Handbook/
84. Marcus B, Tuncel S (1990) Entropy at a weight-per-symbol and embeddings of Markov chains. Invent Math 102:235–266
85. Marcus B, Tuncel S (1991) The weight-per-symbol polytope and scaffolds of invariants associated with Markov chains. Ergod Theory Dynam Syst 11:129–180
86. Marcus B, Tuncel S (1993) Matrices of polynomials, positivity, and finite equivalence of Markov chains. J Amer Math Soc 6:131–147
87. Markley N, Paul M (1981) Matrix subshifts for $\mathbf{Z}^\nu$ symbolic dynamics. Proc Lond Math Soc 43:251–272
88. Markley N, Paul M (1981) Maximal measures and entropy for $\mathbf{Z}^\nu$ subshifts of finite type. In: Devaney R, Nitecki Z (eds) Classical mechanics and dynamical systems. Dekker Notes 70:135–157
89. Meester R, Steif J (2001) Higher-dimensional subshifts of finite type, factor maps and measures of maximal entropy. Pac Math J 200:497–510
90. Mouat R, Tuncel S (2002) Constructing finitary isomorphisms with finite expected coding time. Isr J Math 132:359–372
91. Morse M (1921) Recurrent geodesics on a surface of negative curvature. Trans Amer Math Soc 22:84–100
92. Morse M, Hedlund GA (1938) Symbolic dynamics. Amer J Math 60:815–866
93. Morse M, Hedlund GA (1940) Symbolic dynamics II, Sturmian trajectories. Amer J Math 62:1–42
94. Mozes S (1989) Tilings, substitutions and the dynamical systems generated by them. J Anal Math 53:139–186
95. Mozes S (1992) A zero entropy, mixing of all orders tiling system. In: Walters P (ed) Symbolic dynamics and its applications. Contemp Math 135:319–326
96. Nagy Z, Zeger K (2000) Capacity bounds for the three-dimensional (0, 1) run length limited channel. Trans IEEE Inf Theory 46:1030–1033
97. Nasu M (1986) Topological conjugacy for sofic systems. Ergod Theory Dynam Syst 6:265–280
98. Ornstein D (1970) Bernoulli shifts with the same entropy are isomorphic. Adv Math 4:337–352
99. Parry W (1964) Intrinsic Markov chains. Trans Amer Math Soc 112:55–66
100. Parry W (1977) A finitary classification of topological Markov chains and sofic systems. Bull Lond Math Soc 9:86–92
101. Parry W (1979) Finitary isomorphisms with finite expected code-lengths. Bull Lond Math Soc 11:170–176
102. Parry W (1991) Notes on coding problems for finite state processes. Bull Lond Math Soc 23:1–33
103. Parry W, Schmidt K (1984) Natural coefficients and invariants for Markov shifts. Invent Math 76:15–32
104. Parry W, Tuncel S (1981) On the classification of Markov chains by finite equivalence. Ergod Theory Dynam Syst 1:303–335
105. Parry W, Tuncel S (1982) Classification problems in ergodic theory. In: LMS Lecture Notes, vol 67. Cambridge University Press, Cambridge
106. Pavlov R (2007) Perturbations of multi-dimensional shifts of finite type. Preprint
107. Petersen K (1989) Ergodic theory. Cambridge University Press, Cambridge
108. Quas A, Sahin A (2003) Entropy gaps and locally maximal entropy in Z^d-subshifts. Ergod Theory Dynam Syst 23: 1227–1245
109. Quas A, Trow P (2000) Subshifts of multidimensional shifts of finite type. Ergod Theory Dynam Syst 20:859–874
110. Radin C (1996) Miles of tiles. In: Pollicott M, Schmidt K (eds) Ergodic Theory of Z^d-actions. LMS Lecture Notes, vol 228. Cambridge University Press, Cambridge, pp 237–258
111. Robinson RM (1971) Undecidability and nonperiodicity for tilings of the plane. Invent Math 12:177–209
112. Robinson EA (2004) Symbolic dynamics and tilings of $\mathbf{R}^d$. In: Williams S (ed) Symbolic dynamics and its applications. Proc Symp Appl Math 60:81–120
113. Rudolph D (1990) Fundamentals of measurable dynamics. Oxford University Press, Oxford

114. Schmidt K (1984) Invariants for finitary isomorphisms with finite expected code lengths. Invent Math 76:33–40
115. Schmidt K (1990) Algebraic ideas in ergodic theory. AMS-CBMS Reg Conf 76
116. Schmidt K (1995) Dynamical systems of algebraic origin. Birkhauser, Basel
117. Schwartz M, Bruck S (2008) Constrained codeds as networks of relations. IEEE Trans Inf Theory 54:2179–2195
118. Seneta E (1980) Non-negative matrices and Markov chains, 2nd edn. Springer, Berlin
119. Shannon C (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423,623–656
120. Sinai YG (1968) Markov partitions and C-diffeomorphisms. Funct Anal Appl 2:64–89
121. Smale S (1967) Differentiable dynamical systems. Bull Amer Math Soc 73:747–817
122. Trachtman A (2007) The road coloring problem. Israel J Math, to appear
123. Tuncel S (1981) Conditional pressure and coding. Isr J Math 39:101–112
124. Tuncel S (1983) A dimension, dimension modules and Markov chains. Proc Lond Math Soc 46:100–116
125. Wagoner J (1992) Classification of subshifts of finite type revisited. In: Walters P (ed) Symbolic dynamics and its applications. Contemp Math 135:423–444
126. Wagoner J (2004) Strong shift equivalence theory. In: Walters P (ed) Symbolic dynamics and its applications. Proc Symp Appl Math 60:121–154
127. Walters P (1982) An introduction to ergodic theory. Springer Grad Text Math 79. Springer, Berlin
128. Walters P (1992) Symbolic dynamics and its applications. In: Walter P (ed) Contemp Math 135. AMS, Providence
129. Ward T (1994) Automorphisms of Z^d-subshifts of finite type. Indag Math 5:495–504
130. Weiss B (1973) Subshifts of finite type and sofic systems. Monats Math 77:462–474
131. Williams RF (1973/74) Classification of subshifts of finite type. Ann Math 98:120–153; Erratum: Ann Math 99:380–381
132. Williams S (2004) Introduction to symbolic dynamics. In: Williams S (ed) Symbolic dynamics and its applications. Proc Symp Appl Math 60:1–12
133. Williams S (2004) Symbolic dynamics and its applications. In: Williams S (ed) Proc Symp Appl Math 60. AMS, Providence

Synchronization Phenomena on Networks

Guanrong Chen[1], Ming Zhao[2], Tao Zhou[2], Bing-Hong Wang[2,3]

[1] Department of Electronic Engineering,
City University of Hong Kong, Hong Kong, China

[2] Department of Modern Physics and Nonlinear Science
Center, University of Science and Technology of China,
Hefei Anhui, China

[3] Institute of Complex Adaptive Systems,
Shanghai Academy of System Science, Shanghai, China

Article Outline

Glossary

Synchronization A problem in time-keeping, requiring the coordination of events to operate a system or a task in unison.

Distance A measure between two nodes, defined as the number of edges connecting them through the shortest paths.

Average distance The mean distance, averaged over all pairs of nodes on the network.

Clustering coefficient The probability that two randomly-selected neighboring nodes of a node are directly connected each other.

Node-degree The number of edges incident from a node.

Random-graph network A type of graph obtained by starting with a set of nodes and then adding edges between them at random.

Small-world network A type of graph in which most nodes are not neighbors of each other, but most nodes can be reached from any other node by a small number of connection steps; thus, a small-world network is highly clustered like a regular graph, and yet with a small average distance, just like a random graph.

Scale-free network A type of graph in which a small number of nodes have a large number of connections while a large number of nodes have a small number of connections, whose node-degree distribution typically follows a power-law form, with both structure and dynamics being independent of the network size.

Node-betweenness A measure of the extent to which a given node is occupied by the amount of information passing through it via shortest paths between other nodes, namely, the portion of shortest paths between all pairs of nodes which have data traffic going through this particular node in the network.

Definition of the Subject

The subject under consideration is synchronization on complex networks, with respect to the phenomena and particularly the ability of achieving synchrony of a network of dynamical systems. The subject of synchronization is quite old, but it is a significant one continuously calling for serious and systematic investigation. Ever since the careful study of two synchronous pendulum clocks by the great Dutch scientist Christian Huygens in 1665, the subject has evolved to be an independent and indispensable field of scientific research. The current study of complex networks, on the other hand, is pervading all kinds of sciences, ranging from physical to biological, even to social sciences. Its impact on modern engineering and technology is prominent and will be far-reaching. Typical complex dynamical networks include the Internet, the World Wide Web, various wireless communication networks, metabolic networks, biological neural networks, social relationship networks, financial and economic networks, and so on. As it has turned out today, the study of synchronization phenomena and synchronous behaviors of dynamical systems such as oscillators on complex networks has become overwhelming. This article offers an overview of the state-of-the-art advances and developments of the subject of synchronization on various complex networks, with emphasis on network synchronizability and performance.

Introduction

Many biological, social and technological systems can be properly described by complex networks with nodes representing individuals or organizations and edges characterizing the interactions among them [1,2,3,4,5]. One of the goals in the current studies on complex networks is to understand and explain how the topological properties of a network affect the behaviors of dynamical systems built upon the network. Typical examples include understanding how the topology of the Internet affects the spread of the computer viruses [6,7,8,9,10], how the structure of a power grid affects the cascading failures over time [11,12,13,14,15], how the connecting patterns of an intercommunication network affect its data traffic and dynamics [16,17,18,19,20], and so on.

Synchronous behaviors have been observed in various complex networks in nature and human society [23,24,25,26], and they have been studied for hundreds of years since the systematic investigation of pendulum synchrony by the great Dutch scientist Christian Huygens in 1665 [27].

To understand how network structure affects the synchronizability of a network not only has broad theoretical interest [28], but also has important practical value [29]. One typical case in point is the synchronicity of sensors in biological neural networks, where neurons communicate with each other through synaptic junctions for which a mechanism called asynchronous release is important [30]. There are many careful studies about collective synchronization in the earlier literature, with a basic assumption that dynamical systems of coupled oscillators evolve either on regular networks [31,33,34] or on random networks [35,36]. However, the structures of most real-world networks are neither completely regular nor completely random, but rather, somewhere in between. Thus, it becomes important and even necessary to consider how network structure affects the synchronization process and the synchronizability of the dynamical systems on such networks. Recently, it has been found that networks with small-world effects and scale-free properties are quite different from, and oftentimes achieve synchronization more easily than, regular networks such as lattices [37,38,39,40,41,42,43,44].

The study of synchronization on complex networks has gone through several stages in the past decade, encompassing several important aspects of the subject: various synchronization phenomena on complex networks and their stability analysis, the relationships between structural ingredients and a network's synchronizability, the enhancement or reduction of network synchronizability, etc. The first two are quite well understood today while the last one will be further addressed in this article. First, some basic concepts about synchronization of networked dynamical systems and the associated stability analysis are introduced. Second, some intrinsic relations between network structure and synchronizability are discussed. Third, three types of methods, namely, regulating coupling patterns, modifying network structures, and designing output functions, are introduced for enhancing network synchronizability. Finally, some open questions are posed which are deemed significant for further studies of the important subject of complex network synchronization.

To proceed, some notations are introduced [5], among which three are most significant with respect to network synchronization: average distance L, clustering coefficient C, node-degree k_i of node i, and the corresponding probability density function of degree distribution $p(k)$.

In the past few years, by taking advantage of both high-speed computing power and the huge amount of real data available on the web, scientists were able to search and find some common statistical characteristics shared by many real-world networks. It is found that most real networks have a very small average distance, scaled approximately as $L \sim \ln N$, where N is the size of the network (i. e., the total number of its nodes), while their clustering coefficient is rather large, as compared with *random-graph networks*

(Erdös and Rényi [28]). A network having both of these two characteristics is referred to as a *small-world network*, described by Watts and Strogatz [21]. Moreover, the degree distributions of many real networks obey a power-law form $p(k) \sim k^{-\gamma}$, where $p(k)$ is the probability density function for the corresponding degree distribution, and γ is the power-law exponent (typically $2 < \gamma < 3$) [1,2,3]. The power-law distribution falls off much more gradually than an exponential one, allowing for a few nodes with very large degrees to exist. Networks with power-law degree distributions usually belong to the class of *scale-free networks*, characterized by Barabási and Albert [22].

Basic Concepts of Network Synchronization

A general model of coupled identical oscillators on a network can be described by [40,41]

$$\dot{\boldsymbol{x}}_i = \boldsymbol{F}(\boldsymbol{x}_i) - \sigma \sum_{j=1}^{N} G_{ij} \boldsymbol{H}(\boldsymbol{x}_j) , \quad i = 1, \ldots, N , \qquad (1)$$

where $\dot{\boldsymbol{x}}_i = \boldsymbol{F}(\boldsymbol{x}_i)$ governs the dynamics of the ith oscillator, with state vector $\boldsymbol{x}_i$; $\boldsymbol{H}(\boldsymbol{x}_j)$ is the output function; σ is the coupling strength; $G = [G_{ij}]$ is an $N \times N$ coupling matrix determined by the given coupling pattern among the N oscillators.

In the typical situation when the oscillators are symmetrically coupled, the coupling matrix G has the same form as the graph Laplacian L, i. e., $G = L$, with

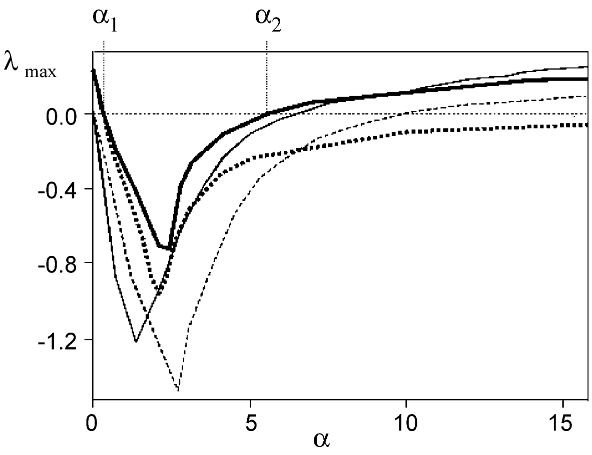
$$L_{ij} = \begin{cases} k_i & \text{for } i = j \\ -1 & \text{for } j \in \Lambda_i \\ 0 & \text{otherwise ,} \end{cases} \qquad (2)$$

where k_i is the degree of node i and Λ_i is the set of its neighboring nodes. In this setting, L is symmetrical and semi-positive definite, and all the rows of L have a zero sum, so that its smallest eigenvalue λ_1 is always a single zero and all the other eigenvalues are strictly positive. Thus, the eigenvalues of L can be ranked as

$$0 = \lambda_1 < \lambda_2 \leq \lambda_3 \leq \cdots \leq \lambda_N .$$

For network (1), the synchronization manifold is an invariant manifold: $\boldsymbol{x}_1 = \boldsymbol{x}_2 = \cdots = \boldsymbol{x}_N = \boldsymbol{s}$, typically satisfies $\dot{\boldsymbol{s}} = \boldsymbol{F}(\boldsymbol{s})$ in engineering applications.

For a dynamical system, the so-called master stability function is usually defined to be the ratio of the largest Lyapunov exponent versus a connectivity parameter of the system [42,45]. For some dynamical systems, the master stability function is negative when $\lambda_2 > \alpha_1/\sigma$ for some constant α_1. In this case, the largest Lyapunov exponent is negative, and consequently the network is synchronizable; moreover, the larger the λ_2 is, the better the network synchronizability will be [40,41].



Synchronization Phenomena on Networks, Figure 1
Four typical master stability functions for coupled Rössler oscillators: chaotic (*bold curve*) and periodic (*regular curve*); with *y*-coupling (*dashed curve*) and *x*-coupling (*dotted curve*). The *vertical ordinate* shows the change of the largest Lyapunov exponent. Curves are all scaled for clearer visualization (after [42])

For some other dynamical systems, the master stability function is negative only within a finite interval (α_1, α_2) [46], over which the largest Lyapunov exponent is negative [42,45], where α_1 and α_2 are constants. In this case, the network is synchronizable for some σ when the eigenratio $R = \lambda_N/\lambda_2$ satisfies $R < \alpha_2/\alpha_1$; moreover, a smaller R indicates a better network synchronizability.

The former case corresponds to networks for which the synchronized region is unbounded (the bold-dashed curve in Fig. 1), and the latter, bounded (the bold-solid curve and the two regular lines in Fig. 1) [42]. In both cases, the right-hand side of the above two inequalities depends only on the dynamics of each individual oscillator and the output function of the network, while the eigenvalue λ_2 and eigenratio R depend only on the Laplacian L. Therefore, the problem of synchronization can be divided into two parts: choosing suitable dynamics (including the aforementioned parameters and output function) and analyzing the eigenvalues of the Laplacian. In fact, these two cases can co-exist [32,101].

The same stability analysis can also be applied to some more complicated coupling patterns [40,41,42,45,47], including the case where G is non-diagonalizable (see Fig. 2 and [48]).

Network (1) has only identical oscillators, while in the real world parameter mismatch between oscillators is very common, so that both the amplitudes and phases of different oscillators become different. However, quite often, only the frequencies of oscillations are of concern in some applications, while the amplitudes are not important. In such cases, phase synchronization is the topic for study, for which the *Kuramoto model* [49,50,51,52,53] is a representative platform.

In the Kuramoto model, oscillators run at arbitrary frequencies and they are coupled through a periodic (e. g., sine) function of their phase differences. More precisely, the model consists of a population of N coupled phase-oscillators $\theta_i(t)$ having natural frequencies ω_i distributed with a given probability density $g(\omega)$, governed by

$$\dot{\theta}_i = \omega_i - \sigma \sum_{j=1}^{N} G_{ij} \sin(\theta_i - \theta_j)\,, \quad i = 1, \ldots, N\,. \tag{3}$$

To measure the synchronization phenomena, an order parameter M is introduced:

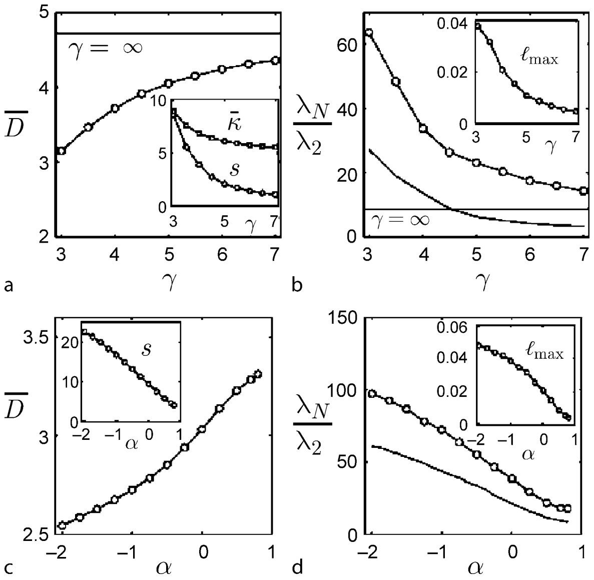
$$M \equiv \left[\left\langle \left| N^{-1} \textstyle\sum_{j=1}^{N} \mathrm{e}^{i\phi_j} \right| \right\rangle \right]\,, \tag{4}$$

where ϕ is a function of θ, and $\langle\cdot\rangle$ and $[\cdot]$ denote the average over time and over different configurations, respectively.

Initially, each node is assigned a random phase. Without coupling, all the oscillators run independently and, at any time, the phases of the oscillators are distributed almost uniformly on the interval $[0, 2\pi]$, yielding $M = O(1/\sqrt{N})$. In this situation, the oscillators are generally not synchronized. With coupling, as the coupling strength gradually increases to beyond a certain threshold, interactions among oscillators become stronger and more inter-influential, which gradually dominate the individual self-oscillations. Eventually, collective synchronization of all oscillators emerges spontaneously. During this transition process, the order parameter M increases from 0 to 1.

Synchronizability Versus Structure

Previous studies have demonstrated that both scale-free and small-world networks are much easier to synchronize than regular lattices [37,38,39,40,41,42,43,44]. At this point, a natural question arises: what makes them easier to synchronize? An intuitive answer might be their average distance, which is much shorter than that of a regular network with the same size. However, after some systematic investigations on the relation between structural ingredients and the network synchronizability, Nishikawa et al. [54] found that as the network becomes more heterogeneous, i. e., the degree distribution becomes wider, a network can become less synchronizable even though its average distance becomes much shorter. Figure 2 gives two examples of this phenomenon. In a semi-random model [55], with the increase of the power-law exponent γ, which makes the network more homogeneous, the network average distance $\bar{D}$ becomes longer and the standard deviation of the degree distribution reduces (Fig. 2a and inset); meanwhile, the eigenratio λ_N/λ_2 of its Laplacian becomes smaller (Fig. 2b), indicating improvement of the network synchronizability. In a growing model of scale-free networks with aging nodes [56], it is also observed that as the average distance increases and the degree distribution becomes more homogeneous, the network gains a better synchronizability (Fig. 2c,d).

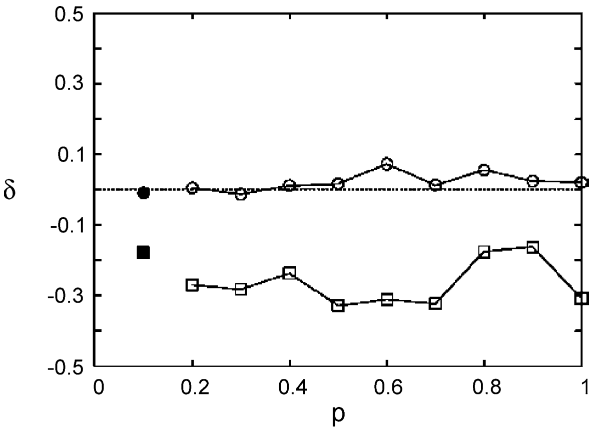


Synchronization Phenomena on Networks, Figure 2

Synchronization of scale-free networks. a, b the semi-random model; c, d the growing model with aging of nodes. The small *insets* are the responses of the indicated parameters with respect to the changing parameters γ or α under the same conditions (after [54])

A heuristic exploration may be given: in a network with a heterogeneous degree distribution, a few "central" oscillators, which interact with a large number of other oscillators, tend to be overloaded by the traffic passing through them. When too many independent traffic signals with different phases and frequencies are traversing through a node at the same time, they cause congestion, leading to the reduction of network synchronizability. The same also happens to overloaded edges [54].

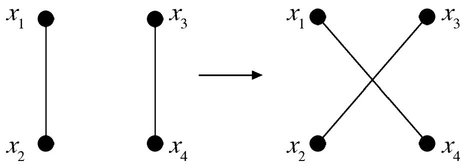
On the other hand, based on experience with WS small-world networks, Hong et al. [57] concluded that the



Synchronization Phenomena on Networks, Figure 3
Behavior of the difference δ of the eigenratio in a WS network with rewiring probability p (after [57])

maximal node betweenness [58,59,60] is a good indicator for network synchronizability: the smaller, the better, and vice versa. To confirm their observation, they calculated the difference of the eigenratio before and after the removal of a node from a WS network [21]. Figure 3 plots the difference $\delta \equiv (\lambda_N/\lambda_2)_{\text{after}} - (\lambda_N/\lambda_2)_{\text{befor}}$. The reduction of the ratio is brought about by the removal of the node with the maximal betweenness (empty squares in the figure). In comparison, random removal of a node makes the eigenratio almost unchanged (empty circles in the figure). This implies that the node with the maximal betweenness plays an important role in determining the synchronizability of the network. However, for scale-free networks, this "maximal betweenness indicator" may not work, as pointed out in [61] with a counterexample given in [62].

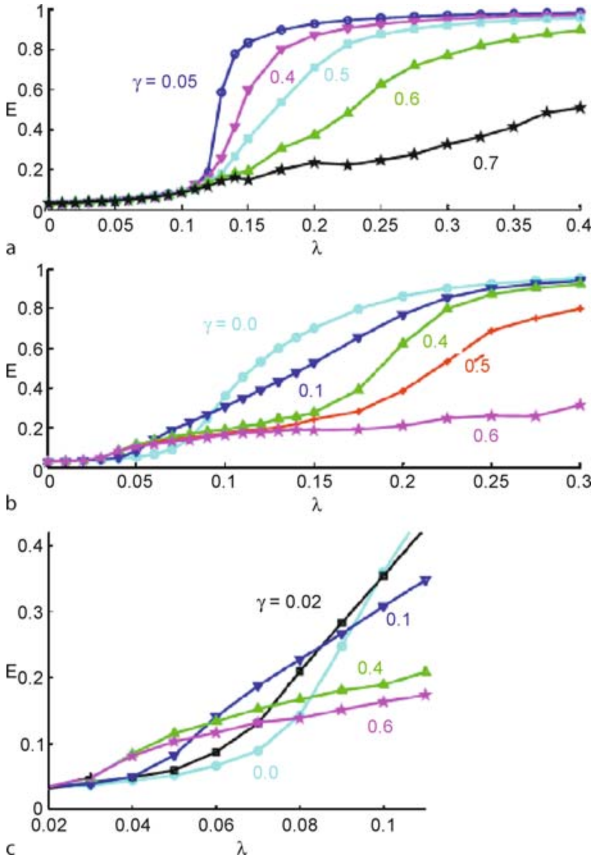
In the above studies, a network is usually modified in order to see how the synchronizability changes as the network structure varies. It is worth emphasizing that during the modification process all the topological ingredients [5] have been changed at the same time, therefore it is impossible to obtain any accurate relation between one particular ingredient and the network synchronizability. Knowing this problem, by using the edge-exchange operation [63,64], Zhao et al. [62] derived some fairly accurate relations between the synchronizability and the average distance as well as the heterogeneity of the degree distribution, on small-world and scale-free network models. Figure 4 presents a sketch of maps of their random interchanging algorithms. The algorithmic operations will change only the network average distance while keeping the degree of each node unchanged. Thus, the relations between the two concerned ingredients can be investigated separately. Extensive simulations have verified that either shortening the average distance or lowering the heterogeneity may lead to a better synchronizability, but only their combination can always ensure that the network will synchronize easily.



Synchronization Phenomena on Networks, Figure 4
Sketch of maps of the random interchanging algorithm (after [62])

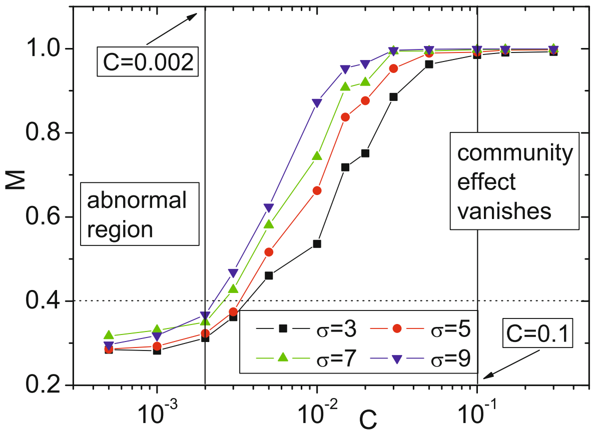
McGraw and Menzinger [65] investigated the relations between the clustering coefficient and network synchronizability, and concluded that for both random-graph and scale-free networks, increasing the clustering coefficient hinders global synchronization if the coupling strength is strong, but it promotes the synchronization of scale-free networks when the coupling strength is weak. Figure 5 shows this phenomenon. The main reason is that the clusters around the hub-nodes promote the formation of frequency-synchronized clusters, but they will inhibit the synchronization of the network as a whole. The early hub synchronization accounts for the slightly enhanced order parameter when the coupling is weak [65,66]. This analysis is based on non-identical oscillators in the Kuramoto model. On the other hand, by means of master stability analysis, Wu et al. [67] reported a negative correlation between the clustering coefficient and synchronizability through a scale-free network model with a tunable clustering coefficient [68].

Besides the main focus on small-world effects and scale-free properties, as described by the clustering coefficient, average distance and degree distribution, some further studies on the effects of other topological ingredients on network synchronization have also been reported, particularly the degree-degree correlation. A network is said to show *assortative* (or *disassortative*) *mixing*, if the nodes having many connections tend to connect to other nodes with many (or few) connections. The extent of this degree-degree correlation can be measured by the Pearson coefficient [69]: its positive (or negative) value indicates assortative (or disassortative) mixing. Di Bernardo et al. found that disassortative networks generally have a better synchronizability than the assortative ones [70,71]. However, later works [72,73] show that the degree distribution, coupling pattern, and degree-degree correlation among the nodes compete with each other in an intrinsic manner,



Synchronization Phenomena on Networks, Figure 5

Order parameter M vs coupling strength λ, for different values of the clustering coefficient γ. a Poisson degree distribution. b, c Power-law degree distribution. c A close-up of the transition region, showing that increase of the clustering coefficient leads to an advanced (lower-λ) transition (after [65])



Synchronization Phenomena on Networks, Figure 6

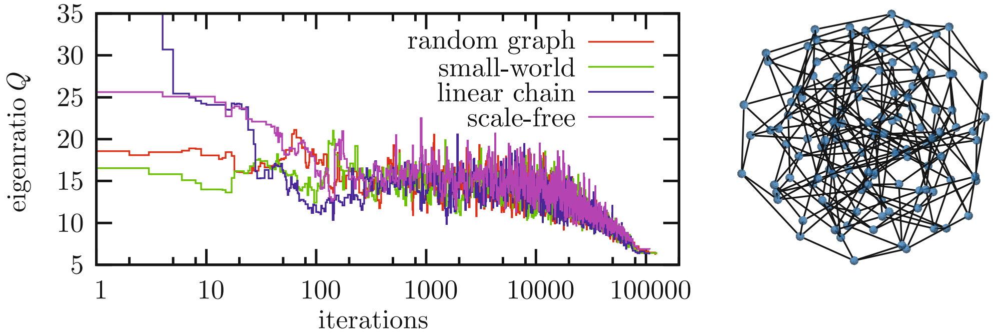
Order parameter M vs. community strength C for different values of the coupling strength σ (after [76])

thereby together determining the network synchronizability. That is, for one coupling pattern, disassortative mixing may predict better synchronizability, while for another coupling pattern, the result can be the opposite.

As we gain more knowledge of various network structures, more attention is paid to the effects of local structures of complex networks on their global behaviors and dynamics. Huang et al. [74] found that in complex networks with prominent clusters, the synchronizability is determined by the interplay between intercluster and intracluster edges: a network is mostly synchronizable when the numbers of the two types of edges are approximately equal. If not equal, for example as the number of intracluster edges increases, an abnormal synchronization phenomenon appears: although the network average distance becomes smaller, the network synchrony is weakened or even destroyed.

Furthermore, the synchronization phenomenon of a complex network with a community structure has also been discussed. Qualitatively, a *community* is defined as a subset of nodes within a network with the property that the connections among the nodes therein are denser than those within the other parts of the network [75]. Zhou et al. studied phase synchronization in a network with a community structure [76]. Defining the edges connecting two nodes in one community as *internal edges*, and those connecting nodes between two communities as *external edges*, the ratio of the number of external edges to the number of internal edges can be used to characterize the strength of the community structure, denoted by C. Clearly, a smaller C corresponds to sparser external edges thus a more prominent community structure. Figure 6 shows the relationship between the order parameter M and the community strength C for different coupling strengths σ. It is found from Fig. 6 that a strong community structure will hinder global synchronization no matter what the coupling strength is, but this effect will vanish when the fraction of external connections exceeds 0.1.

Using a modified simulated annealing algorithm, Donetti et al. [77] generated an entangled network with optimal synchronizability. These kinds of networks are shown to have an extremely homogeneous structure: distributions of node degrees, distances, betweenness, and loops are all very uniform. Also, these networks are characterized by short average distances and large loops, with no well-defined community structures. In the approach of [77], rewiring is applied, i. e., at each time step, the number of rewiring trials is randomly extracted from an exponential distribution. Except for rewiring which reduces or increases the eigenratio, and except for operations that



Synchronization Phenomena on Networks, Figure 7
Eigenratio *Q* as a function of the number of algorithmic iterations. Starting from different initial configurations, all the networks are converted via iterations to *entangled networks* (after [77])

disconnect the network, for different initial configurations the optimization process will always lead to the same optimal result. Figure 7 shows the changes of the eigenratio in the optimizing process and the resultant network configuration.

Enhancing Network Synchronizability

With a clearer understanding of the relations between the network structure and synchronizability, a natural question about how to enhance the network synchronizability is in order. Some effective synchronizability-enhancement methods are introduced in this section.

Coupling Pattern Regulation

In general, scale-free networks are much harder to synchronize than random networks with the same size and the same average degree. One reason is that in scale-free networks, there are some "central" oscillators that interact with a large number of other nodes [54]. Thus, when too many independent signals with different phases and frequencies are traversing through a "central" oscillator at the same time they may have conflicts, thereby causing traffic congestion. Hence, generally speaking, the more heterogeneous the degree distribution, the more difficult for the network to synchronize. It is also known that in scale-free networks, when the oscillators are coupled symmetrically, oscillators with larger degrees usually approach the final synchronized state first, and then the others with smaller degrees synchronize to them gradually [65]. Therefore, when the oscillators are coupled asymmetrically, if the coupling strength from the "central" oscillators to the other nodes are stronger than the reverse, the network will synchronize much easier and faster.

Based on this idea, Motter, Zhou and Kurths [78,79,80] proposed a new coupling pattern, which we will call the MZK pattern, which can sharply improve network synchronizability. After that, quite a few methods for regulating coupling patterns are brought forward to improve network synchronizability, some static and some dynamic.

Static Coupling Patterns

In static coupling patterns, the elements of the coupling matrix are formulated based on the MZK pattern, as

$$G_{ij} = L_{ij}/k_i^{\beta} \,, \tag{5}$$

where β is a tunable parameter. The coupling is weighted when $\beta \neq 0$, and unweighted when $\beta = 0$. In spite of the asymmetry of this coupling matrix G, it can be proved that all the eigenvalues of G are nonnegative reals with only one eigenvalue being zero if the network is connected. Rewrite Eq. (5) as

$$G = D^{-\beta} L \,, \tag{6}$$

where $D = \operatorname{diag}(k_1, \dots, k_N)$ is a diagonal matrix and L is the Laplacian. From the identity

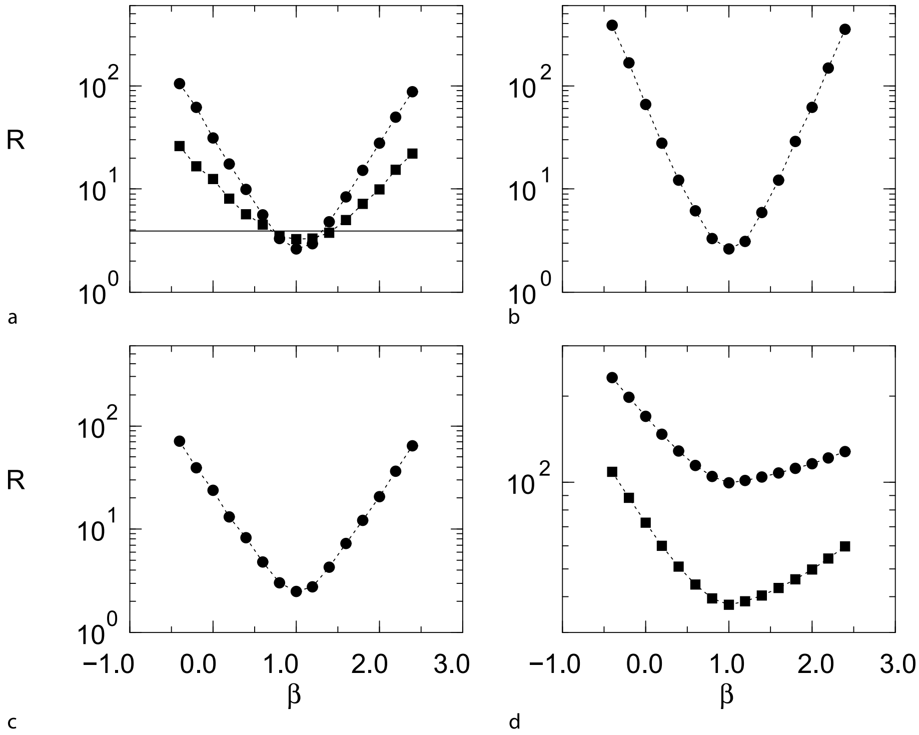
$$\det(D^{-\beta} L - \lambda I) = \det(D^{-\beta/2} L D^{-\beta/2} - \lambda I) \,, \tag{7}$$

where I is the $N \times N$ identity matrix, one can prove that the spectrum of G is the same as that of the following symmetric matrix:

$$H = D^{-\beta/2} L D^{-\beta/2} \,. \tag{8}$$

Similarly to the case of matrix G, if the network is connected then all eigenvalues of H other than the single $\lambda_1 = 0$, are positive. With $\beta = 1$, the matrix H is a normalized Laplacian. Thus, if the network is connected and $N \geq 2$, then

$$0 < \lambda_2 \leq N/(N-1) \,, \quad 2 \leq \lambda_N \leq N/(N-1) \,. \tag{9}$$



Synchronization Phenomena on Networks, Figure 8

Eigenratio *R* as a function of β for four kinds of complex networks specified in [78]. For each model, the synchronizability peaks at $\beta = 1.0$ (after [78])

Figure 8 shows the changes of the eigenratio R with the parameter β in four kinds of complex networks specified in [78]. It can be seen that the eigenratio R has a well-defined minimum at $\beta = 1$ in all cases. Mathematically, this means that the best results are obtained when the matrix D has a square-root. It is also clear that the more heterogeneous the network is, the more prominent the minimum of the eigenratio R becomes.

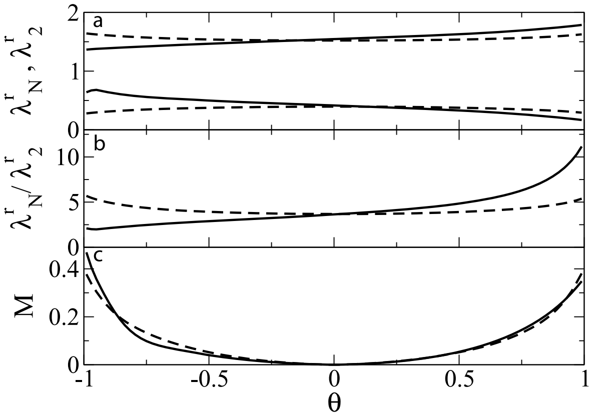
By explicitly relating the asymmetry in the connections to an age order among different nodes, Hwang et al. [81] found that age-ordered networks provide a better propensity for synchronization. The main reason is that an older node becomes weaker, therefore more easily influenced by other nodes. In this coupling pattern, the off-diagonal entries of the zero-row-sum coupling matrix G are

$$G_{ij} = -a_{ij} \frac{\Theta_{ij}}{\sum_{j \in \Lambda_i} \Theta_{ij}} , \tag{10}$$

where a_{ij} are the elements of the adjacency matrix A ($a_{ij} = 1$ if nodes i and j are connected, and $a_{ij} = 0$ otherwise), and $\Theta_{ij} = (1 - \theta)/2$ (or $\Theta_{ij} = (1 + \theta)/2$) for $i > j$ (or $i < j$). The parameter $\theta \in (-1, 1)$ governs the coupling asymmetry in the network: the limit $\theta \to -1$ (or $\theta \to 1$) gives a unidirectional coupling, where the old (or young) nodes drive the young (or old) ones. When $\theta = 0$, the coupling pattern degenerates to the MZK pattern at $\beta = 1$.

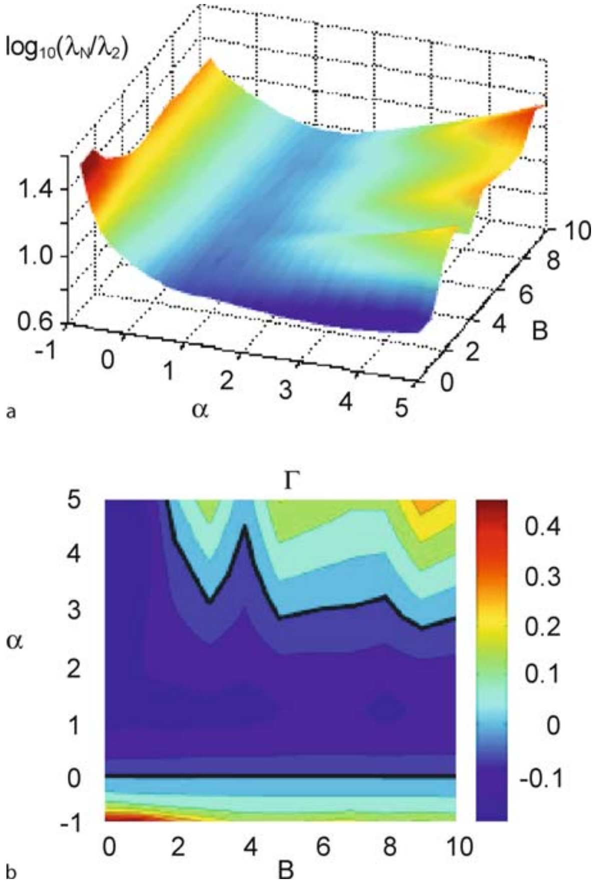
For a generic θ, the spectrum of the coupling matrix G is in the complex plane and the complex eigenvalues appear in pairs of complex conjugates ($\lambda_1 = 0$; $\lambda_\ell = \lambda_\ell^r + j\lambda_\ell^i$, $\ell = 2, \ldots, N$). It can be proved that (i) $0 < \lambda_2^r \leq \cdots \leq \lambda_N^r \leq 2$, and (ii) $|\lambda_\ell^i| \leq 1, \forall \ell$. The best propensity for synchronization is then ensured when both the ratio $\lambda_N^r / \lambda_\ell^r$ and $M \equiv \max_\ell \{|\lambda_\ell^i|\}$ are simultaneously made as small as possible.

In scale-free network models, the age of a node can be denoted by the time when it is being added to the network. The class of scale-free networks under study is generated from the Barabasí–Albert model [82,83]. For comparison, a highly homogeneous random network with an arbitrary initial age ordering is considered, with the average degree being equal to that of the scale-free network. Figure 9 shows the variation of the synchronizability of the two networks versus the parameter θ. For the random network, symmetric coupling makes the ratio $\lambda_N^r / \lambda_2^r$ smallest, while for the scale-free model, the propensity for synchronization is better (or worse) when $\theta \to -1$ (or $\theta \to 1$). As for M, there are only very small differences between



Synchronization Phenomena on Networks, Figure 9

a λ^r_N **and** λ^r_2**, b** λ^r_N/λ^r_2**, and c** M**, vs.** θ**, for a scale-free network (**$m = 5$ **and** $B = 0$**,** ***solid curves*****) and random network (*****dashed curves*****) (after [81])**



Synchronization Phenomena on Networks, Figure 10

a λ_N/λ_2 **for scale-free networks vs. the parameter space** (α, B)**. b** $\Gamma = \log(\lambda_N/\lambda_2) - [\log(\lambda_N/\lambda_2)]_{\alpha=0}$ **vs.** (α, B)**. The domain with** $\Gamma < 0$ **is outlined by the** ***black contours*** **drawn (after [85])**

the scale-free and the random-network models. Thus, it is concluded that in scale-free networks, the network synchronizability is enhanced when the dominant coupling direction is from older to younger nodes [84].

Taking the edge-weights into account, Chavez et al. [85,86] investigated the propensity for synchronization of some weighted complex networks, where the weight in an arbitrary edge, ℓ_{ij}, is defined as its traffic load [87], which quantifies the traffic of shortest paths which make use of that edge. In this coupling pattern, the off-diagonal entries of the zero-row-sum coupling matrix G are

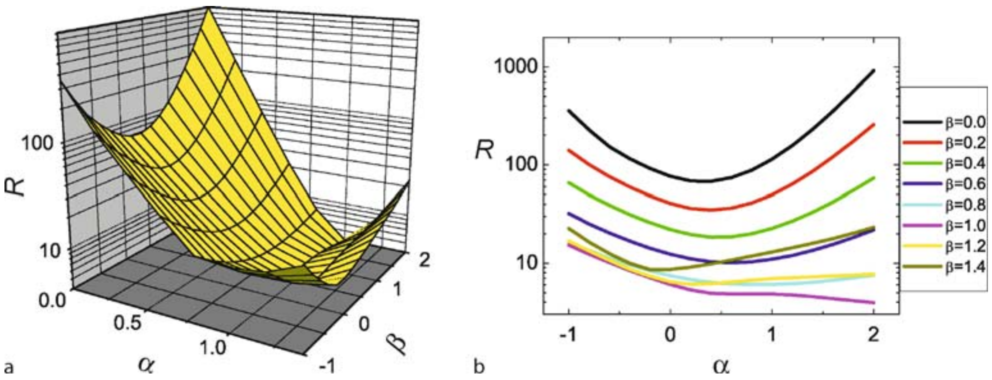
$$G_{ij} = -\frac{\ell^{\alpha}_{ij}}{\sum_{j\in\Lambda_i}}\,\ell^{\alpha}_{ij}\,, \tag{11}$$

where α is a tunable parameter, and ℓ_{ij} is the load of the edge connecting nodes i and j.

Although G is asymmetric for all α, just like the MZK pattern, it can be proved that all its eigenvalues are nonnegative reals with only one zero eigenvalue if the network is connected. The case of $\alpha = 0$ corresponds to the best synchronizability condition for the MZK pattern. From Eq. (11), it can be seen that in the limit of $\alpha = +\infty$ (or $\alpha = -\infty$) only the edges with the largest (or smallest) loads ℓ_{ij} are selected as the incoming edges for each node i. Therefore, this generates a network with at least N directed edges, which can be either connected or disconnected. In the connected (or disconnected) case, the ratio λ_N/λ_2 will be equal to 2 (or $+\infty$), thus yielding a very strong (or weak) condition for synchronization.

Figure 10a shows the logarithm of λ_N/λ_2 in the parameter space (α, B) for the above-discussed model [82,83]. Parameter B is used to regulate the heterogeneity of the degree distribution. It can be observed that the surface of λ_N/λ_2 has a prominent minimum when $\alpha \simeq 1$ for all values of B above a given threshold $B_c > 0$, which means that the weighting procedure based on edge loads always enhances the network propensity for synchronization. The quantity $\Gamma = \log(\lambda_N/\lambda_2) - [\log(\lambda_N/\lambda_2)]_{\alpha=0}$ shown in Fig. 10b may be used to quantify the synchronizability enhancement.

The coupling patterns proposed by both Hwang et al. [81] and Chavez et al. [85,86] can enhance the propensity for network synchronization. The former works well only for age-ordered networks, while the latter requires the knowledge of the load on each edge of the whole network. Therefore, a general coupling pattern using only local information would be very desirable. Based on the idea that different nodes should play different roles in a network, Zhao et al. [73] proposed a coupling pattern which



Synchronization Phenomena on Networks, Figure 11
a Eigenratio R in the parameter plane (α, β). b R vs. α for different values of parameter β (after [73])

requires only the degrees of neighboring nodes. The coupling matrix G of this pattern is given by

$$G_{ij} = \begin{cases} -k_j^{\alpha}/S_i^{\beta} & \text{for } j \in \Lambda_i \\ S_i/S_i^{\beta} & \text{for } i = j \\ 0 & \text{otherwise}\,, \end{cases} \tag{12}$$

where $S_i = \sum_{j \in \Lambda_i} k_j^{\alpha}$. When $\alpha = \beta = 0$, this coupling pattern degenerates to the symmetric coupling pattern [45], where the case of $\alpha = 0$ corresponds to the MZK pattern [78] and the case of $\beta = 1$ is equivalent to the one introduced in [80] (see Eq. (15) in [80] for more details). Although this G is asymmetric for all α with $\beta \neq 0$, it can also be proved that all its eigenvalues are non-negative reals with only one zero eigenvalue, if the network is connected. Figure 11 shows some simulation results. From the figure, it can be concluded that there is always some parameter region in which the eigenratio R is smaller than that of the symmetrically coupled case ($\alpha = \beta = 0$) and that of the optimal case with the MZK pattern ($\alpha = 0$ and $\beta = 1$).

From the viewpoint of gradient fields, Wang et al. [88] also derived a coupling pattern that has the same configuration as Eq. (12) with $\beta = 1$.

Dynamic Coupling Patterns

The coupling patterns discussed above are all based on a network having a fixed structure which remains unchanged throughout the synchronizing process.

Zhou et al. [89] investigated synchronization in a scale-free network of chaotic oscillators, where the coupling strength of a node develops adaptively according to the local synchronizing property between the node and its neighbors. In this coupling pattern, the off-diagonal entries of the zero-row-sum coupling matrix G are

$$G_{ij} = -a_{ij} W_{ij}\,, \tag{13}$$

where $W_{ij} > 0$ is the coupling strength from node j to node i if they are connected. Here, suppose that the strength between node i and all its k_i neighbors increases uniformly among the k_i connections, in order to suppress its difference Δ_i from the mean activity of its neighbors; namely,

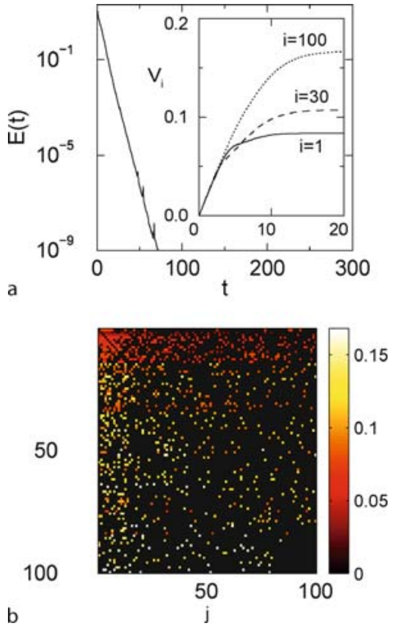
$$G_{ij}(t) = -a_{ij} V_i(t)\,, \quad \dot{V}_i = \gamma \Delta_i/(1 + \Delta_i)\,, \tag{14}$$

where $\Delta_i = |\boldsymbol{H}(\boldsymbol{x}_i) - (1/k_i) \sum_j a_{ij} \boldsymbol{H}(\boldsymbol{x}_j)|$, and $\gamma > 0$ is the adaptation parameter. It is clear that, in this adaptive coupling scheme, the input weight ($W_{ij} = V_i$) and the output weight ($W_{ji} = V_j$) of node i are generally asymmetrical.

Next, synchronization of a network of coupled Rössler oscillators and a chaotic foodweb model on Barabasí–Albert scale-free networks are considered, and two cases of unbounded and bounded stability zones are investigated, respectively. When the stability zone is unbounded, the transition to synchronization is shown in Fig. 12a. Starting from random initial conditions on the chaotic attractors, the local synchronization difference $\Delta \gg 1$, and the input weights of each node, both increase uniformly on the whole network, i. e., $W_{ij} = V_i(t) \approx \gamma t$ (Fig. 12a, inset). After a short period of time, the weights V_i of different nodes develop at different rates and then converge to different values $\tilde{V}_i$. The input weight is smaller on average for nodes with larger degrees k_i (Fig. 12b). Here, the synchronization error is measured by averaging all local errors over the nodes: $E(t) = \langle |\boldsymbol{x}_i - \langle \boldsymbol{x}_i \rangle| \rangle$.

The dependence of the input weight of a node on its degree follows a power law,

$$V(k) \sim k^{-\theta}\,, \tag{15}$$



Synchronization Phenomena on Networks, Figure 12

a Transition to synchronization in an adaptive network of Rössler oscillators, indicated by the (averaged) synchronization error $E(t) = \langle |x_i - \langle x_i \rangle | \rangle$. *Inset*: the input strength $V_i(t)$ vs. time over three nodes. b The weighted coupling matrix $\tilde{G}$ crystallized after the adaptation (for the foodweb model) (after [89])

with exponent $\theta = 0.48 \pm 0.01$ for both oscillator models. Importantly, this scaling is also robust to the variation of network parameters, such as the minimal degree M (Fig. 13b), which should not be confused with the order parameter M elsewhere, the system size N (Fig. 13c), and the orders of magnitudes of the adaptation parameter γ (Fig. 13d).

When the stability zone is bounded, synchronization can always be achieved by the adaption mechanism of Eq. (14) if $\gamma \leq \gamma_c$ for a threshold γ_c somewhat depending on N and the oscillator dynamics. The two resulting weighted networks display the same power-law behavior as in Eq. (15), but with different exponents: $\theta = 0.54 \pm 0.01$ (Rössler oscillator) or $\theta = 0.36 \pm 0.01$ (foodweb). The eigenratio R for the weighted networks, after the adaptation, and for the unweighted networks (symmetric coupling) is calculated as a function of N (Fig. 14a), and as a function of the ratio $S_{\max}/S_{\min}$ (Fig. 14b), where $S_{\max}$ and $S_{\min}$ are the maximum and minimum intensities of the variable coupling strengths of the model. Clearly, this adaptive coupling scheme is more effective than symmetric coupling for network synchronization.

Huang [90] investigated another adaptive coupling pattern, in which a node is coupled with its neighbors non-uniformly through different coupling strengths, and showed that they have better synchronizability than other networks with symmetric coupling patterns.

In all the coupling patterns discussed above, whether the coupling pattern is static or dynamic, only the coupling strength is tunable while the connectivity matrix always remains unchanged. However, as is intuitively clear, network synchronizability can also be significantly improved by evolving the graph topology giving rise to a time-varying connectivity matrix. This has been recently confirmed by Boccaletti et al. [91].

It has been shown [91] that to make a network synchronizable, either the coupling matrix $G(t) = G$ remains unchanged, or if starting from an initial wiring condition $G(0) = G_0$, the coupling matrix $G(t)$ commutes at any time with G_0, i. e., $G_0 G(t) = G(t) G_0, \forall t$. At any time, a zero-row-sum symmetric commuting matrix $G(t)$ can be constructed, as

$$G(t) = V \Lambda(t) V^{\mathrm{T}} , \tag{16}$$

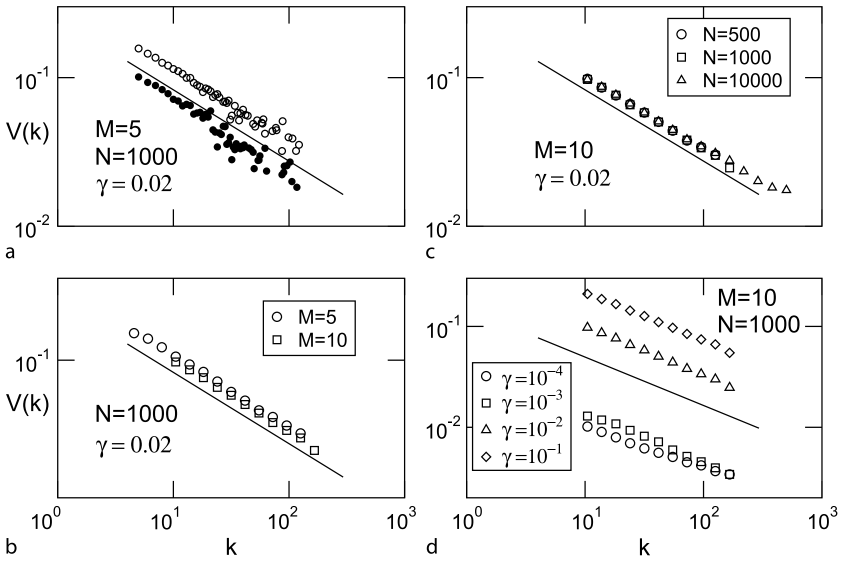
where $V = \{\boldsymbol{v}_i, \ldots, \boldsymbol{v}_N\}$ is an orthogonal matrix with columns being the eigenvectors of G_0, and $\Lambda(t) = \mathrm{diag}[0, \lambda_2(t), \ldots, \lambda_N(t)]$ with $\lambda_i(t) > 0, \forall i > 1$. This set of matrices is referred to as the dissipative commuting set of $G(0)$. A condition to ensure the network synchronization will be stable is

$$S_i = \lim_{T \to \infty} \frac{1}{T} \int_0^T \Lambda_{\max}(\sigma \lambda_i(t')) \mathrm{d}t' < 0 \quad \forall i \neq 1 , \tag{17}$$

where $\Lambda_{\max}(\sigma \lambda_i)$ is the maximal transversal (conditional) Lyapunov exponent along the direction of the ith eigenvector, and S_i is its time average. Hence, it does not require $\Lambda_{\max}(\sigma \lambda_i(t)) < 0$ at all times. One can even construct a commutative evolution such that at each time there exists one eigenvalue λ_i for which $\Lambda_{\max}(\sigma \lambda_i(t)) > 0$, and yet obtain a stable synchronization manifold. Thus, interestingly, synchronization in a dynamical network can be achieved even in the case where each individual commutative graph does not give rise to synchronized behavior.

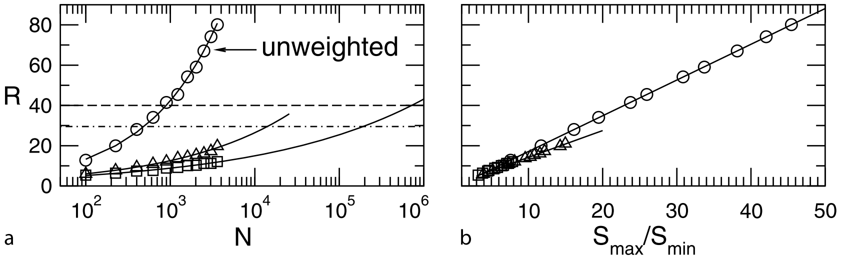
Modifications of Network Structures

It is well known that the synchronizability of a dynamical network is determined simultaneously by the network coupling pattern, the dynamical characteristics of the oscillators on its nodes, and the network structure. In the above, several cases with variable coupling patterns have been discussed. For some real-world networks, however, the coupling pattern cannot be modified at will. Thus, if the dynamics of the oscillators are given and fixed, and



Synchronization Phenomena on Networks, Figure 13

Average input weight $V(k)$ of nodes with degree k as a function of k for a network of Rössler oscillators (*empty circles*) and the food-web model (*filled circles*) (a), and its dependence on various parameters, M (b), N (c), and γ (d), where the M should not be confused with the order parameter M elsewhere (after [89])



Synchronization Phenomena on Networks, Figure 14

Eigenratio R as a function of N (a), and S_{max}/S_{min} (b). The networks are synchronizable if $R < R_\epsilon$ in a, Rössler oscillators (*squares*), $R_\epsilon = 40$ (*dashed curve*), foodweb model (*triangles*), $R_\epsilon = 29$ (*dashed-dotted curve*) (after [89])

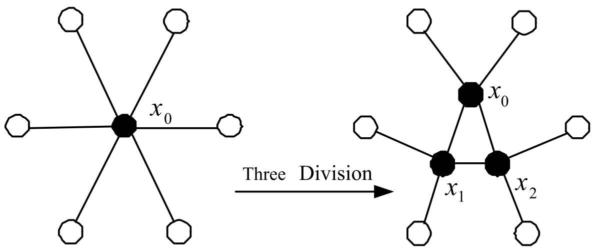
if the coupling patterns cannot be changed, then the only way to enhance the network synchronizability is to make a change to the network structure.

There are some effective techniques to enhance the network synchronizability by modifying the network structure, as further discussed below in the rest of this section.

Reducing Maximal Betweenness

In scale-free networks, the average distance is often very short while the node-degree and node-betweenness distributions are both quite broad. The bottleneck for the network synchronizability seems to be the maximal node betweenness [57]. In order to reduce the node betweenness of the hubs, Zhao et al. [92] suggested a method of structural perturbations. Specifically, for a hub x_0, $m - 1$ auxiliary nodes, labeled as $x_1, \dots, x_{m-1}$, are added around it. These m nodes are fully connected together. Then, all the edges incident from x_0 are re-distributed to all the nodes x_i (including x_0 itself), $i = 0, 1, \dots, m - 1$. After this process, the betweenness of x_0 is divided into m almost equal parts associating with these m nodes. This process is called m-division. A sketch map of a 3-division process on node x_0 is shown in Fig. 15.

Due to the huge sizes of many real-life networks, it is usually impossible to obtain the node betweenness from a complex network. Fortunately, studies have shown that there exists a strongly positive correlation between the node-degree and the node-betweenness in Barabasí–Albert networks and some other heterogeneous networks [87,93]. That is, a node with larger degree has

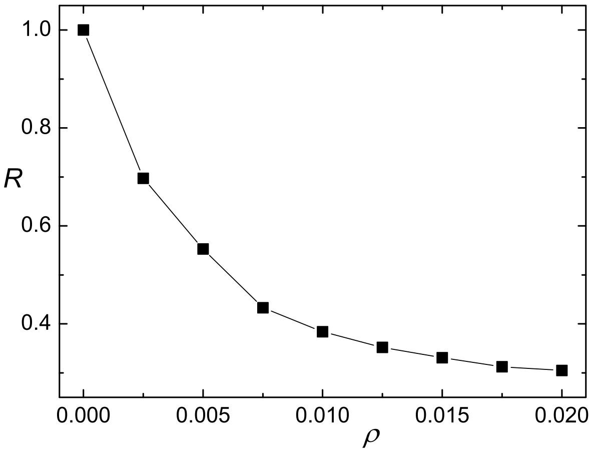


Synchronization Phenomena on Networks, Figure 15

Sketch map for the 3-division process on x_0. The *solid circle* on the *left* is the node x_0 with degree 6. After the 3-division process, this x_0 is divided into 3 nodes, x_0, x_1 and x_2, which are fully connected. The six edges incident from x_0 are then re-distributed to all the three nodes (after [92])

higher node-betweenness statistically. Therefore, for practical reasons, it can be assumed that nodes with higher betweenness are those with larger degrees in Barabasí–Albert networks.

To further explore how the structural perturbations affect the network synchronizability, the eigenratios before and after the m-division process were compared in [92] for a Barabasí–Albert scale-free network with the coupling matrix being Laplacian. For use in the rest of the article, we define a characteristic value $R = r'/r$, in which r and r' are the eigenratios before and after the division, respectively. Figure 16 shows the correlation between R and the probability ρ of the divided nodes. It is clear that even the m-division of a tiny fraction of nodes can sharply enhance network synchronizability.



Synchronization Phenomena on Networks, Figure 16

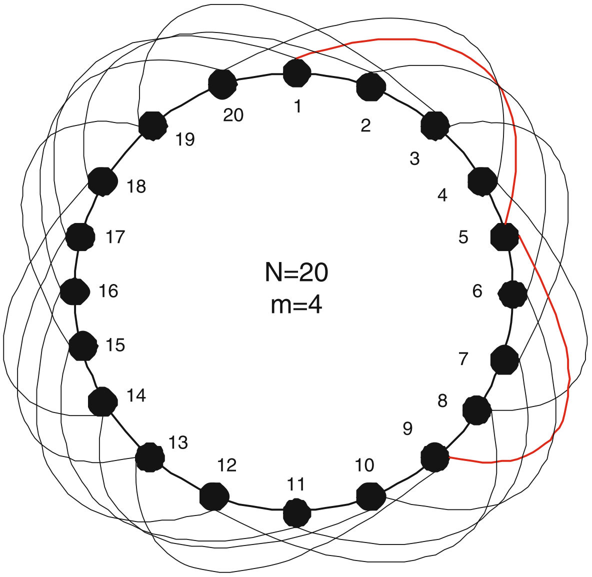
Behavior of value R vs. the fraction of divided nodes ρ. As the number of divided nodes increases, R is reduced, leading to better synchronization (after [92])

Shortening the Average Distance

Zhou et al. [94] investigated the synchronizability of a network model named *crossed double cycles* (CDCs). They not only clarified the relationship between average distance and network synchronizability, but also provided a possible way to make a network more synchronizable.

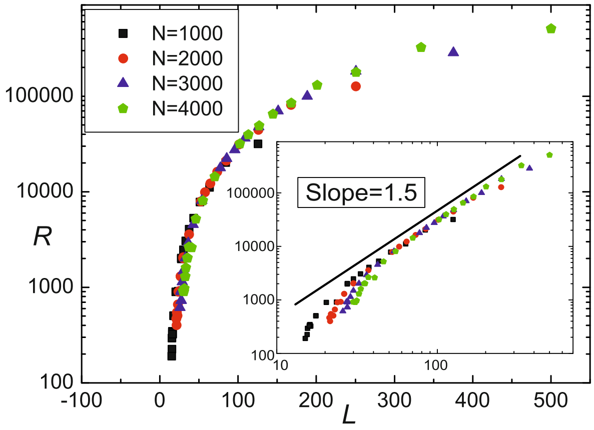
In the language of graph theory [95,96,97], a cycle C_N denotes a network consisting of N nodes (vertices) x_1, $\dots, x_N$. These N nodes are arranged in a ring, and the nearest two nodes are connected to each other. Thus, C_N has N edges connecting the nodes $x_1x_2, x_2x_3, \dots,$ $x_{N-1}x_N, x_Nx_1$. The set of all such CDCs, denoted by $G(N, m)$, can be constructed by adding two edges, called crossed edges, to each node in C_N. The two nodes connecting by a crossed edge have distance m within C_N. For example, the network $G(N, 3)$ can be constructed from C_N by connecting $x_1x_4, x_2x_5, \dots, x_{N-1}x_2, x_Nx_3$ together. A sketch map of $G(20, 4)$ is shown in Fig. 17 for illustration.

Figure 18 shows how the average distance L affects the network synchronizability (measured by the characteristic value R). It is clear that the network synchronizability is very sensitive to the average distance: as L increases, R sharply spans more than three magnitudes. And the network synchronizability is remarkably enhanced by reducing L. When the crossed length m is not too small or too large (compared to N), networks with the same average distance have approximately the same synchronizability,



Synchronization Phenomena on Networks, Figure 17

Sketch map of $G(20, 4)$ (after [94])



Synchronization Phenomena on Networks, Figure 18
Characteristic value *R* vs. average distance *L* of CDCs. The *black squares*, *red circles*, *blue triangles* and *green pentagons* represent the cases of $N = 1000$, 2000, 3000 and 4000, respectively. The *inset* shows the same data in log-log plot, indicating that the characteristic value *R* approximately obeys a power-law form $R \sim L^{1.5}$. The *solid line* has slope 1.5, for comparison (after [94])

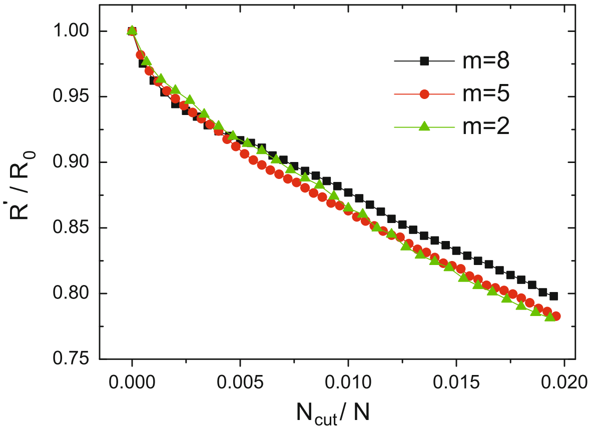
regardless of the network sizes. More interestingly, the numerical results show that the characteristic value R approximately obeys a power-law form, as $R \sim L^{1.5}$ (inset of Fig. 18).

Decoupling Nodes by Removing Heavily-Loaded Edges

In the synchronization process, not only hubs may be the bottlenecks but some edges with large loads may also limit the network synchronizability. Yin et al. [99] found that a scale-free network can become more synchronizable after some of its heavily-loaded edges have been removed. To reduce the computational cost, they used local information to approximately rank the edges, according to the values of $k_i \times k_j$, where i and j denote two adjacent nodes connected by an edge. Subsequently, at each time step, an edge with the highest rank is removed, i. e., the two nodes are decoupled at both sides of their connecting heavily-loaded edge. After this operation, the characteristic value is decreased, as shown by Fig. 19.

Designing the Output Function

Very recently, the relationship between graph theory and network synchronizability received some special attention [100]. For example, Duan et al. [101,102,103] found that for networks with disconnected complementary graphs, adding edges will often increase their synchronizability. The complementary graph of a given graph G is defined to be the graph consisting of all the nodes of G and all the edges that are not in G.

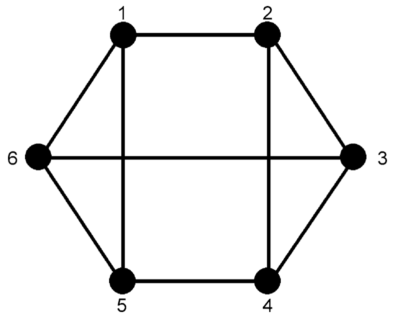


Synchronization Phenomena on Networks, Figure 19
Changes of the synchronizability as a function of the proportion of cut edges N_{cut}/N for different values of the average distance (after [99])

In addition, they found [101,102] that when the couplings between nodes are symmetric, an unbounded synchronized region is always easier to analyze than a bounded synchronized region (see Sect. "Basic Concepts of Network Synchronization" to recall their definitions). Therefore, to effectively enhance network synchronizability, they presented a design method for the output function (i. e., $\boldsymbol{H}$ in network (1), or the inner linking matrix in the linear coupling case), such that the resultant network has an unbounded synchronized region, for the case where the synchronous state is an equilibrium of the network.

If the synchronous state is an equilibrium, then both $D\boldsymbol{F}(\boldsymbol{s}(t))$ and $D\boldsymbol{H}(\boldsymbol{s}(t))$ in network (1), as discussed in Sect. "Enhancing Network Synchronizability" (part B), reduce to constant matrices, denoted by F and H, respectively. The synchronized region is the stability region of the matric pencil $F + \alpha H$ with respect to parameter α. It can be proved that there exists a matrix H of rank 1 (meaning that only one component in each state vector is used for coupling), such that the stability region is unbounded. The method for obtaining the desired output function is outlined below: first, take a column vector b such that (F, b) is stabilizable [104]; then, find a matrix $P = P^{\mathrm{T}}$ such that $FP + PF^{\mathrm{T}} - 2bb^{\mathrm{T}} < 0$; consequently, taking $k = b^{\mathrm{T}} P^{-1}$ leads to the stability of $F - \alpha bk$ for all α in the unbounded region; finally, $H = bk$ is the matrix to be found.

For illustration, synchronization of a simple 6-node network (shown in Fig. 20) is studied, where each node



Synchronization Phenomena on Networks, Figure 20
A network of 6 nodes (after [101])

is located with a third-order smooth Chua's circuit [105]. At first, arbitrarily take the output function

$$H = \begin{pmatrix} 0.8348 & 9.6619 & 2.6591 \\ 0.1002 & 0.0694 & 0.1005 \\ -0.3254 & -8.5837 & -0.9042 \end{pmatrix} . \qquad (18)$$

But the network does not synchronize. The states of node 1 are shown in Fig. 21a. Then, let $b = (0, 0, 1)^{\mathrm{T}}$ and $k = (0.0708, -0.15590, 0.4296)$, and then set $H = bk$, so synchronization is achieved as guaranteed by the theory. Figure 21b shows that the states of node 1 quickly reach the equilibrium.

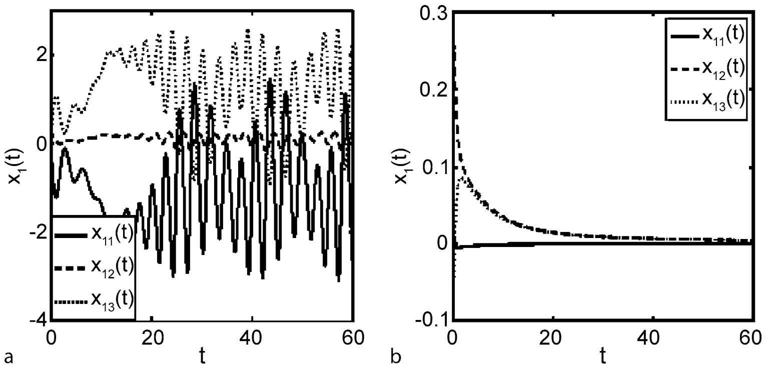
Future Research Outlook

Complex network synchronization is a rapidly growing subject attracting increasing attention from various fields of physics, engineering, mathematics, and biology alike. Despite the current great advances and progress, there are still many important open questions.

In the studies of static coupling, Nishikawa and Motter [48] once pointed out that optimal global synchronizability, with eigenratio being equal to 1, can be obtained from a directed network structure without loops. Even if adding one loop of length 2 (in a directed network, two opposite edges between node *i* and node *j* can be considered as a loop of length 2), the eigenratio will be doubled [73,85]. Another scenario is shown by extending the conclusion in [48] to the case of non-identical oscillators [106]. Some further works in this direction will be helpful for in-depth understanding about the role of loops in network synchronization.

In the studies of dynamic coupling, the cost of coupling has not been taken into account. However, cost is usually very significant in some self-driven systems (for example, in wireless sensor networks [107,108] and in distributed autonomous robotic systems [109]). For each node to report its current state to the neighbors (or to detect the states of all its neighbors) requires a certain amount of power, while the total power assigned to each node is often limited, even if such communications are possible. Yet, as found in collective behaviors of biological swarms, a few effective leaders can well organize the whole population [110]. And a recent study has pointed out that partial coupling is more than enough to keep the coherence of self-propelled particles [111]. Therefore, it is very natural to expect to synchronize a complex network with a very low cost, which is an important issue for further investigation.

Very recently, there are some attempts at detecting the network structures with the help of the synchronization phenomenon on complex networks [112,113,114]: to discover the hierarchical community structure by the dynamic time scales of the network synchronization process [112,113], or to infer the complete connectivity of a network from its stable response dynamics [114], etc. These seem quite useful for optimal network design, analy-



Synchronization Phenomena on Networks, Figure 21
States of node 1 (after [101])

sis, and utilization in general, therefore should be pursued with special efforts.

Similar to the aforementioned open questions, many theoretically attractive and practically important problems about various aspects of synchronization on complex networks can be posted and described. As the network research further evolves in different fields, many new dynamical phenomena and analytic issues will also emerge. Importance notwithstanding, the subject of "Synchronization Phenomenon of Networks" will continue to prove itself an theoretically interesting and technically challenging subject for scientific research in the years to come.

Acknowledgments

The authors thank the American Physics Society and the Elsevier Publisher for their permission to use some of their published simulation figures, which have all been referenced (in the figure captions). The authors also thank Dr. Jian-Guo Liu for his helpful comments and suggestions. G.R.C. was supported by the NSFC/RGC joint research scheme under grant N-CityU107/07. B.H.W. was supported by the NNSFC under Grants 10472116 and A0524701, and the Specialized Program under the Presidential Funds of CAS. T.Z. was supported by NNSFC under Grants 10635040 and 70471033.

Bibliography

1. Albert R, Barabási AL (2002) Rev Mod Phys 74:47
2. Dorogovtsev SN, Mendes JFF (2002) Adv Phys 51:1079
3. Newman MEJ (2003) SIAM Rev 45:167
4. Boccaletti S, Latora V, Moreno Y, Chaves M, Hwang DU (2006) Phys Rep 424:175
5. da F Costa L, Rodrigues FA, Travieso G, Boas PRV (2007) Adv Phys 56:167
6. Pastor-Satorras R, Vespignani A (2001) Phys Rev Lett 86:3200
7. Zhu CP, Xiong SJ, Tian YJ, Li N, Jiang KS (2004) Phys Rev Lett 92:218702
8. Zhou T, Yan G, Wang BH (2005) Phys Rev E 71:046141
9. Zhou T, Fu ZQ, Wang BH (2006) Prog Nat Sci 16:452
10. Zhou T, Liu JG, Bai WJ, Chen GR, Wang BH (2006) Phys Rev E 74:056109
11. Motter AE, Lai YC (2002) Phys Rev E 66:065102
12. Goh KI, Lee DS, Kahng B, Kim D (2003) Phys Rev Lett 91:148701
13. Motter AE (2004) Phys Rev Lett 93:098701
14. Zhou T, Wang BH (2005) Chin Phys Lett 22:1072
15. Galstyan A, Cohen P (2007) Phys Rev E 75:036109
16. Guimerá R, Díaz-Guilera A, Vega-Redondo F, Cabrales A, Arenas A (2002) Phys Rev Lett 89:248701
17. Tadić B, Thurner S, Rodgers GJ (2004) Phys Rev E 69:036102
18. Zhao L, Lai YC, Park K, Ye N (2005) Phys Rev E 71:026125
19. Yan G, Zhou T, Hu B, Fu ZQ, Wang BH (2006) Phys Rev E 73:046108
20. Wang BH, Zhou T (2007) J Korean Phys Soc 50:134
21. Watts DJ, Strogatz SH (1998) Nature 393:440
22. Barabási AL, Albert R (1999) Science 286:509
23. Strogatz SH, Stewart I (1993) Sci Am 269:102
24. Gray CM (1994) J Comput Neurosci 1:11
25. Néda Z, Ravasz E, Vicsek T, Barabási AL (2000) Phys Rev E 61:6987
26. Glass L (2001) Nature 410:277
27. Chen G, Dong X (1998) From Chaos Order. World Scientific, Singapore
28. Erdös P, Rényi A (1960) Pub Math Ins Hung Acad Sci 5:17
29. Wang XF (2002) Int J Bifurc Chaos 12:885
30. Heidelberger R (2007) Nature 450:625
31. Heagy JF, Carroll TL, Pecora LM (1994) Phys Rev E 50:1874
32. Stefański A, Perlikowski P, Kapitaniak T (2007) Phys Rev E 75:016210
33. Wu CW, Chua LO (1995) IEEE Trans Circuits Syst I 42:430
34. Jost J, Joy MP (2001) Phys Rev E 65:016201
35. Gade PM (1996) Phys Rev E 54:64
36. Manrubia SC, Mikhailov AS (1999) Phys Rev E 60:1579
37. Gade PM, Hu CK (1999) Phys Rev E 60:4966
38. Gade PM, Hu CK (2000) Phys Rev E 62:6409
39. Lago-Fernández LF, Huerta R, Corbacho F, Sigüenza JA (2000) Phys Rev Lett 84:2758
40. Wang XF, Chen G (2002) Int J Bifurc Chaos Appl Sci Eng 12:187
41. Wang XF, Chen G (2002) IEEE Trans Circuits Syst I 49:54
42. Barahona M, Pecora LM (2002) Phys Rev Lett 89:054101
43. Jiang PQ, Wang BH, Bu SL, Xia QH, Luo XS (2004) Int J Mod Phys B 18:2674
44. Lind PG, Gallas JAC, Herrmann HJ (2004) Phys Rev E 70:056207
45. Pecora LM, Carrol TL (1998) Phys Rev Lett 80:2109
46. Hu G, Yang J, Liu W (1998) Phys Rev E 58:4440
47. Pecora LM, Barahona M (2005) Chaos Complex Lett 1:61
48. Nishikawa T, Motter AE (2006) Phys Rev E 73:065106
49. Kuramoto Y (1975) In: Araki H (ed) Internaltional Symposium on Mathematical Problems in Theoretical Physics. Lecture Notes in Physics, vol 30. Springer, New York
50. Kuramoto Y (1984) Chemical Oscillations, Wave and Turbulence. Springer, Berlin
51. Kuramoto Y, Nishikawa I (1987) J Stat Phys 49:569
52. Pikovsky A (2001) Synchronization. Cambridge University Press, Cambridge
53. Acebrón JA, Bonilla LL, Vicente CJP, Ritort F, Spigler R (2005) Rev Mod Phys 77:137
54. Nishikawa T, Motter AE, Lai YC, Hoppensteadt FC (2003) Phys Rev Lett 91:014101
55. Newman MEJ, Strogatz SH, Watts DJ (2001) Phys Rev E 64:026118
56. Dorogovtsev SN, Mendes JFF (2000) Phys Rev E 62:1842
57. Hong H, Kim BJ, Choi MY, Park H (2004) Phys Rev E 69:067105
58. Freeman L (1979) Soc Netw 1:215
59. Newman MEJ (2001) Phys Rev E 64:016132
60. Zhou T, Liu JG, Wang BH (2006) Chin Phys Lett 23:2327
61. Fan ZP (2006) Complex Networks: From Topology to Dynamics. Ph D Thesis, City University of Hong Kong
62. Zhao M, Zhou T, Wang BH, Yan G, Yang HJ, Bai WJ (2006) Physica A 371:773
63. Maslv S, Sneppen K (2002) Science 296:910
64. Kim BJ (2004) Phys Rev E 69:045101(R)
65. McGraw PN, Menzinger M (2005) Phys Rev E 72:015101
66. Gómez-Gardeñes J, Moreno Y, Arenas A (2007) Phys Rev Lett 98:034101

67. Wu X, Wang BH, Zhou T, Wang WX, Zhao M, Yang HJ (2006) Chin Phys Lett 23:1046
68. Holme P, Kim BJ (2002) Phys Rev E 65: 026107
69. Newman MEJ (2002) Phys Rev Lett 89:208701
70. di Bernardo M, Garofalo F, Sorrentino F (2007) Int J Bifurc Chaos 17:3499
71. Sorrentino F, di Bernardo M, Cuéllar GH, Boccaletti S (2006) Physica D 224:123
72. Chavez M, Hwang DU, Martinerie J, Boccaletti S (2006) Phys Rev E 74:066107
73. Zhao M, Zhou T, Wang BH, Ou Q, Ren J (2006) Eru Phys J B 53:375
74. Huang L, Park K, Lai YC, Yang L, Yang K (2006) Phys Rev Lett 97:164101
75. Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D (2004) Proc Natl Acad Sci USA 101:2658
76. Zhou T, Zhao M, Chen G, Yan G, Wang BH (2007) Phys Lett A 368:431
77. Donetti L, Hurtado PI, Muñoz MA (2005) Phys Rev Lett 95:188701
78. Motter AE, Zhou C, Kurths J (2005) Phys Rev E 71:016116
79. Motter AE, Zhou C, Kurths J (2005) Europhys Lett 69:334
80. Motter AE, Zhou C, Kurths J (2005) AIP Conf Proc 776:201
81. Hwang DU, Chavez M, Amann A, Boccaletti S (2005) Phys Rev Lett 94:138701
82. Dorogovtsev SN, Mendes JFF, Samukhin AN (2000) Phys Rev Lett 85:4633
83. Krapivsky PL, Redner S (2001) Phys Rev E 63:066123
84. Zou Y, Zhu J, Chen G (2006) Phys Rev E 74:046107
85. Chavez M, Hwang DU, Amann A, Hentschel HGE, Boccaletti S (2005) Phys Rev Lett 94:218701
86. Chavez M, Hwang DU, Amann A, Boccaletti S (2006) Chaos 16:015106
87. Goh KI, Kahng B, Kim D (2001) Phys Rev Lett 87:278701
88. Wang X, Lai YC, Lai CH (2007) Phys Rev E 75:056205
89. Zhou C, Kurths J (2006) Phys Rev Lett 96:164102
90. Huang D (2006) Phys Rev E 74:046208
91. Boccaletti S, Hwang DU, Chavez M, Amann A, Kurths J, Pecora LM (2006) Phys Rev E 74:016102
92. Zhao M, Zhou T, Wang BH, Wang WX (2005) Phys Rev E 72:057102
93. Barthélemy M (2004) Eur Phys J B 38:163
94. Zhou T, Zhao M, Wang BH (2006) Phys Rev E 73:037101
95. Bondy JA, Murty USR (1976) Graph Theory with Applications. MacMillan, London
96. Bollobás B (1998) Modern Graph Theory. Springer, New York
97. Xu JM (2003) Theory and Application of Graphs. Kluwer, Dordrecht
98. Newman MEJ, Watts DJ (1999) Phys Rev E 60:7332
99. Yin CY, Wang WX, Chen G, Wang BH (2006) Phys Rev E 74:047102
100. Comellas F, Gago S (2007) J Phys A Math Theor 40:4483
101. Duan Z, Chen G, Huang L (2007) Phys Rev E 76:056103
102. Duan Z, Chen G, Huang L (2007) Phys Lett A 372:3741
103. Duan Z, Liu C, Chen G (2008) Physica D 237:1006
104. Friedland B (1986) Control System Design. McGraw-Hill, New York
105. Tsuneda A (2005) Int J Bifurc Chaos 15:1
106. Um J, Han SG, Kim BJ, Lee SI (2008) (unpublished)
107. Akyildiz IF, Su W, Sankarasubramaniam Y, Cayirci E (2002) Comput Netw 38:393
108. Oqren P, Fiorelli E, Leonard NE (2004) IEEE Trans Automat Contr 49:1292
109. Arai T, Pagello E, Parker LE (2002) IEEE Trans Robot Automat 18:655
110. Couzin LD, Krause J, Franks NR, Levin SA (2005) Nature 433:513
111. Zhang HT, Chen M, Zhou T (2007) arXiv:0707.3402
112. Arenas A, Díaz-Guilera A, Pérez-Vicente CJ (2006) Phys Rev Lett 96:114102
113. Boccaletti S, Ivanchenko M, Latora V, Pluchino A, Rapisarda A (2007) Phys Rev E 75:045102(R)
114. Timme M (2007) Phys Rev Lett 98:224101

Synergetics: Basic Concepts

HERMANN HAKEN
Institut für Theoretische Physik, Universität Stuttgart, Stuttgart, Germany

Article Outline

Glossary

Synergetics Science of cooperation.

Pattern A pattern is essentially an arrangement. It is characterized by the order of the elements of which it is made rather than by the intrinsic nature of these elements (Norbert Wiener).

Self-organization Formation of spatio-temporal patterns (structures) and/or performance of functions without an “ordering hand”.

State vector Set of time- or time-independent variables that characterize the state of a system.

Evolution equations Determine the temporal evolution of the state vector. May be deterministic, stochastic or both.

Control parameter One or a set of (mostly externally) fixed parameters in the evolution equations.

Spectrum Set of eigenvalues belonging to linear stability equations with boundary conditions.

Stability of a system System returns after a (small) perturbation of its state vector into original state.

Instability Loss of stability.

Order parameters Collective variables that determine the macroscopic behavior of systems.

Slaving principle A general theorem that allows the reduction of the variables of a system to order parameters (close to instability).

Trajectory Smooth curve $q(t)$ of solution of evolution equation in q-space.

Attractor Region in the state vector space ("q-space") to which all neighboring states are attracted in the course of time.

Fixed point, stable Point in q space to which all neighboring trajectories converge in course of time.

Limit cycle, stable A closed trajectory to which all neighboring trajectories converge.

Probability distribution function Function that determines the probability of a random variable r to have fixed value $r = r_0$.

Fokker Planck equation Evolution equation for probability density function, based on drift and diffusion.

Normal form Especially simple polynomial expression that still captures the essential features, e. g. of the right hand side of deterministic evolution equations.

Schrödinger picture of quantum mechanics In it operators are time-independent, while the wave-function ("state vector") is time-dependent and determined by the Schrödinger equation.

Heisenberg picture in quantum mechanics The state vector is time-independent, while the operators are time-dependent and determined by Heisenberg equations of motion.

Fluctuating forces Stochastic (random) forces appearing in evolution equations.

Quantum classical correspondence Establishes relation between quantum mechanical density matrix and classical quasi-probability distribution.

Symmetry Invariance of a system against specific transformations (e. g. mirror symmetry).

Group Set of elements with specific multiplication rules (axioms).

Dynamical system System whose state vector changes in the course of time deterministically.

Langevin equation Originally: evolution equation for velocity of a Brownian particle subject to damping and fluctuating force.

Generalized Langevin equation General evolution equations that contain both a deterministic and a stochastic part ("fluctuating forces").

Hamilton operator Classical Hamilton function, in which variables, e. g. position x and momentum p, are replaced by quantum mechanical operators.

Spatial coordinate (vector *x*) in one, two or three dimensions.

$$\Delta \quad \text{Laplace operator (in 1,2 or 3 dimensions)}\,.$$

$$\nabla \quad \text{Vector} \left(\frac{\mathrm{d}}{\mathrm{d}x_1}, \frac{\mathrm{d}}{\mathrm{d}x_2}, \frac{\mathrm{d}}{\mathrm{d}x_3}\right) \text{ in 1,2 or 3 dimensions}\,.$$

The Role of Synergetics in Science

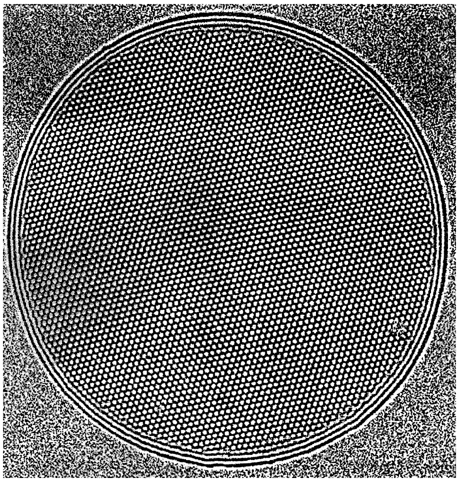
In science, we may essentially distinguish between two trends:

1. The accumulation of knowledge
2. Information reduction in the sense of finding general principles, common features.

In physics, such unifying approaches are well known: the unification of magnetism, electricity and, later on, weak and other interactions leading eventually to a unified field theory. General relativity unifies concepts of space, time and gravitation. While these unifications take place at a fundamental level, one may ask whether it is worthwhile to look also for unifications at say more macroscopic or phenomenological levels. One example is thermodynamics, another the theory of phase transitions of systems in thermal equilibrium by means of the renormalization group approach, or the concept of fractals, etc.

The main goal of Synergetics is the search for unifying principles for systems that are composed of many individual parts or components, and that may show the phenomenon of self-organization, i. e. the spontaneous formation of spatial, temporal, spatial-temporal or functional structures. The systems under discussion are, in the widest sense of the word, open physical systems whose states are maintained by an in- and outflux of energy, matter and /or information. A typical and well known example is that of a fluid in a pan that is uniformly heated from below. When the temperature difference between the lower and upper surface exceeds a critical value, the formerly homogeneous fluid develops roll or hexagonal patterns in which the fluid moves in a specific manner (Fig. 1).

As it turned out, the general principles originally elaborated in physics, can also be applied to many other systems, such as in biology, economy, ecology, sociology,



Synergetics: Basic Concepts, Figure 1
Hexagonal pattern of a fluid (liquid helium) uniformly heated from below [12]

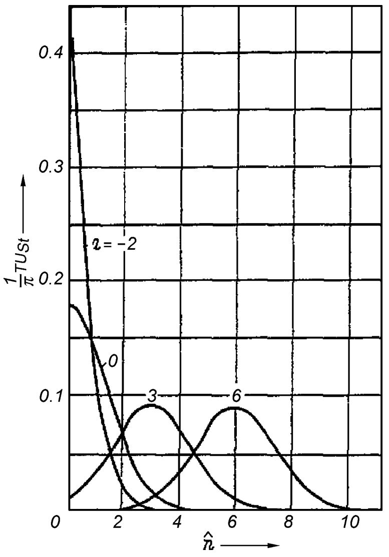
management theory, psychology etc. In spite of the great variety of the individual systems with their components quite different in nature, such principles apply to large classes of phenomena. This is achieved by restricting the study to situations where the systems undergo qualitative changes at macroscopic scales. Here macroscopic means "with time and length scales large compared to those of the individual components".

This leads to the definition of Synergetics as given in the preamble of the Springer Series in Synergetics: "An ever increasing number of scientific disciplines deal with complex systems. These are systems that are composed of many parts which interact with one another in a more or less complicated manner. One of the most striking features of many such systems is their ability to spontaneously form spatial or temporal structures. A great variety of these structures are found, in both the inanimate and the living world. In the inanimate world of physics and chemistry, examples include the growth of crystals, coherent oscillations of laser light, and the spiral structures formed in fluids and chemical reactions. In biology we encounter the growth of plants and animals (morphogenesis) and the evolution of species. In medicine we observe, for instance, the electromagnetic activity of the brain with its pronounced spatio-temporal structures. Psychology deals with characteristic features of human behavior ranging from simple pattern recognition tasks to complex patterns of social behavior. Examples from sociology include the formation of public opinion and cooperation or competition between social groups."

In recent decades, it has become increasingly evident that all these seemingly quite different kinds of structure formation have a number of important features in common. The task of studying analogies as well as differences between structure formation in these different fields has proved to be an ambitious but highly rewarding endeavor. The Springer Series in Synergetics provides a forum for interdisciplinary research and discussions on this fascinating new scientific challenge. It deals with both experimental and theoretical aspects. The scientific community and the interested layman are becoming ever more conscious of concepts such as self-organization, instabilities, deterministic chaos, nonlinearity, dynamical systems, stochastic processes, and complexity. All of these concepts are facets of a field that tackles complex systems, namely Synergetics.

The Laser Paradigm

This example elucidates central concepts used in Synergetics in a qualitative fashion. An example for the laser device (an acronym for light amplification by stimulated emission of radiation, originally called optical maser [121]) is the gas laser in which gas atoms are enclosed in a tube at the end-faces of which mirrors are mounted. The mirrors serve the purpose of reflecting light running in axial direction sufficiently often so that the corresponding light wave stays for an extended period in this device and can interact intensely with the atoms. The atoms are excited from the outside, e. g. by a pump light source. After having been excited, each atom can spontaneously emit a light wave track. In the usual case of a *lamp*, these wave tracks are emitted independently of each other and the amplitudes are Gaussian distributed. When the pump intensity is increased beyond a critical value, the present state gives way to a single wave with *stable amplitude* on which small *amplitude fluctuations* and *phase diffusion* are superimposed [53]. The pump intensity serves as *control parameter*. At its critical value, the old state becomes *unstable*. The emerging coherent wave acts as *order parameter* that via stimulated emission forces the electrons of the gas molecules to emit light waves in a coherent fashion. This action of the order parameter on the individual parts of the system is called *slaving principle*. If the pump power is increased further, more instabilities can appear, and a variety of temporal but also spatio-temporal patterns of light waves may appear, such as laser light chaos [55] or ultrashort laser pulses. The first laser threshold shows the typical features of a phase transition of a system in thermal equilibrium, namely critical slowing down, critical fluctuations and symmetry breaking [25,46,53,58,119], as well as the emergence of a c-number amplitude of the quantized light field (Fig. 2).



Synergetics: Basic Concepts, Figure 2

The stationary distribution function of the laser light intensity as a function of the normalized intensity $\hat{n}$. The individual *curves* refer to different normalized pump power values a, where $a < 0$ below threshold, $a = 0$ at threshold, $a > 0$ above threshold (after [116])

The Hierarchical Structure of Synergetics

Before I discuss the mathematical approach in detail and to provide the ground for farther reaching applications, I hint at the three levels of Synergetics:

1. *The microscopic theory*, based either on microscopic equations, such as in the laser example, those of quantum mechanics and quantum field theory, or in biology on mathematical models on the behavior of individual parts of a system. At this level, concepts, such as order parameters and enslavement (cf. Sect. "The Laser Paradigm"), can be mathematically derived.
2. *Phenomenological Synergetics* directly starts from concepts, such as order parameters and enslavement, which then may be cast into mathematical relations.
3. *Semantic Synergetics* deals with cases where a mathematical formulation is (at present or in principle) not possible, but still formulations using concepts and relationships unearthed in Synergetics are applicable.

A general goal of Synergetics consists in elaborating relationships between levels 1, 2, 3.

In the present article I will mainly focus my attention on the mathematical formulation dealing with 1. and 2.

Basic Equations

The basic equations are classical or quantum mechanical evolution equations, in which the temporal evolution of the microscopic quantities under consideration is described by ordinary or partial differential equations. Since the systems are open, the inputs and outputs of energy, matter and/or information must be taken care of, which, quite often, appears in the form of coupling to heat baths in the sense of thermodynamics. In open systems, these heat baths must be kept at different temperatures, in order to maintain the non-equilibrium state of the system. The heat bath variables can be eliminated which gives rise to differential equations which contain "pumping" and "damping" terms as well as fluctuating (stochastic) forces. In the case of quantum mechanical equations the stochastic forces are operators. With the inclusion of stochastic forces, the classical or quantum mechanical equations acquire the character of stochastic differential equations which may be called "generalized Langevin equations".

Depending on the definition of the random forces, we may distinguish between the $\hat{I}$ to, the Statonovich and the Klimontovich approach [62,72,134]. As is well known in statistical physics, Langevin equations can be converted into equations for distribution functions, such as e. g. the Fokker–Planck equation. A further approach, mainly used in quantum mechanics, but also in models on sociodynamics, is the master equation.

In order not to overload this article, I will focus my attention on the treatment of evolution equations.

This approach seems to be particularly suited for the treatment of phase transition- like phenomena, i. e. the transitions between qualitatively different states of a system. If noise is neglected and transients are not treated, these transitions are called bifurcations [5,22,49,71,79, 84,91].

At the microscopic level the systems are described by a state vector q with components $q_1, \ldots, q_n$ which may also be space dependent, $q_j = q_j(x, t)$, where x is a one, two or three dimensional vector. The time dependence is described by evolution equations of the form of a vector equation.

$$\dot{q} = N(q, \nabla, \alpha) + F(q, \nabla, \alpha) \,. \tag{1}$$

The dot $^{\cdot}$ means time-derivative. N is a vector valued function that depends on q in a nonlinear fashion. ∇ indicates spatial derivatives (of any order) or non-local integrations e. g. of the form

$$\int K(x, x')q(x')\,\mathrm{d}x' \tag{2}$$

where K is a matrix.

α represents a set of fixed control parameters. If not otherwise stated, we explicitly treat only one control parameter. Equation (1) must be supplemented by appropriate boundary and initial conditions. F is a vector valued stochastic function of time with vanishing mean.

Method of Solution

We assume that for a certain control parameter value α_0 the state vector as solution of Eq. (1) is known, $q = q_0$. The following cases have been considered, see e. g. [62]:

a) q_0 is a stable fixed point (Section "Instability of a Fixed Point")
b) q_0 is a stable limit cycle (Section "Instability of a Limit Cycle, $q_0(t)$ [62]")
c) q_0 is a stable n-dimensional torus. (Section "Instability of Tori [62]")

Now the control parameter value is changed and the stability of the system is checked by means of linear stability analysis [52].

Instability of a Fixed Point We first elucidate our general procedure by means of the instability of an originally stable fixed point. This procedure differs from the classical approach of bifurcation theory [90,124] in two important aspects:

1. The role of the fluctuating forces is fully taken into account in order to be able to make contact with the theory of phase transitions in the Landau sense [86].
2. The approach covers the surrounding of the fixed point in order to deal with relaxation processes towards the newly evolving stable states.

The hypothesis

$$q(t) = q_0 + W(t) \tag{3}$$

is inserted into (1) and the Eq. (1) with $F \equiv 0$ linearized with respect to $W(t)$,

$$\dot{W} = LW \tag{4}$$

where L may be a linear differential (or integral) linear operator.

The solutions are of the form

$$W(x,t) = \mathrm{e}^{\lambda_k t} \sum_{d=0}^{D} t^d v_{k,d}(x) \tag{5}$$

where $D > 0$ may happen if the corresponding eigenvalue λ_k is degenerate. In the following we consider $D = 0$ and $v_{k,d} = v_k$. The unstable modes $v_k \equiv v_u$ are connected with

$$\operatorname{Re}\lambda_k \geq 0\,, \tag{6}$$

the stable modes $v_k \equiv v_s$ with

$$\operatorname{Re}\lambda_k < 0\,. \tag{7}$$

It is assumed that $\operatorname{Re}\lambda_k < A < 0$, A fixed, if the eigenvalues are discrete.

We decompose the wanted solution to the original non-linear and stochastic equations into a super position of modes determined by the instability analysis whereby we distinguish between the unstable and stable modes. The amplitudes of the unstable modes are the order parameters. Inserting

$$q(t) = q_0 + \sum_u \xi_u(t) v_u(x) + \sum_s \xi_s(t) v_s(x) \tag{8}$$

into the Eqs. (1) and projecting both sides of the resulting equation on the stable and unstable modes, we obtain equations of the form

$$\dot{\xi}_u = \lambda_u \xi_u + \hat{N}_u\left(\{\xi_u\}, \{\xi_s\}\right) + \hat{F}_u\left(\{\xi_u\}, \{\xi_s\}\right) \tag{9}$$

$$\dot{\xi}_u = \lambda_s \xi_s + \hat{N}_s\left(\{\xi_u\}, \{\xi_s\}\right) + \hat{F}_s\left(\{\xi_u\}, \{\xi_s\}\right)\,. \tag{10}$$

λ_u, λ_s are the eigenvalues (6), (7), which are assumed to be discrete. By a suitable, in general nonlinear, transformation to new variables, $\tilde{N}(\{\xi_u\})$ can be cast into a particularly simple form ("normal form" theory [101,103], initiated by Poincaré [113]).

If the eigenvalues $\lambda_u, 0 > \operatorname{Re}\lambda_s > -|B|$ are a continuous function of an index, e. g. a wave number k, wave packets of $\xi_u(t)$ are used as new order parameter variables Ξ and $\lambda_u(k)$ is replaced by an operator $\Lambda_u(-i\frac{\mathrm{d}}{\mathrm{d}x})$ in one space-dimension or, more generally, $\Lambda_u = (-i\nabla)$ [62]. For a related approach in fluid dynamics cf. [105].

The central idea of further procedure consists in eliminating the amplitudes of the stable modes. This is achieved by the *slaving principle* [56,62,65,148] which allows us to express the amplitudes of the stable modes in terms of the unstable modes

$$\xi_s(t) = f_s\left(\{\xi_u(t)\}, t\right)\,, \tag{11}$$

where ξ_s, ξ_u are taken at the same time t. The explicit time-dependence of f_s stems exclusively from that of the fluctuating forces. f_s can be explicitly calculated in terms of a series expansion in powers of the order parameters. For practical purposes, in general only a few terms are needed.

For a general discussion of the convergence of this series see [62]. When noise is neglected, contact can be made with center manifold theory [76,112], which originally was a mere existence theory and was not constructive. For more recent developments, see books on bifurcation theory. A related approach is based on time-scale separation: The slowly damped or undamped modes serve as order parameters, which enslave the rapidly damped modes. A special case is adiabatic elimination.

Resulting Langevin Equations The enslaved mode amplitudes can be expressed by the order parameters and inserted in (9), so that closed equations for the order parameters alone result.

$$\dot{\xi_u} = \lambda_u \xi_u + \tilde{N}_u\left(\{\xi_u\}\right) + \tilde{F}_u\left(\{\xi_u\}, t\right) \tag{12}$$

where $\tilde{N}$ is a polynominal of $\xi(x, t)$ starting with at least second order. $\tilde{F}$ is a stochastic force. A simple, yet prototypical example is (with a single order parameter $\xi = \xi_u$)

$$\dot{\xi} = \lambda\xi + a\xi^2 - b\xi^3 + F(t)\,, \quad b > 0 \tag{13}$$

or

$$\dot{\xi} = -\frac{\partial V(\xi)}{\partial \xi} + F(t)\,, \tag{14}$$

with the potential

$$V = -\frac{\lambda}{2}\xi^2 - \frac{a}{3}\xi^3 + \frac{b}{4}\xi^4\,. \tag{15}$$

If λ_u, λ_s (6), (7) represent a continuous spectrum, (generalized) Ginzburg–Landau equations result [62]. For example, the complex Ginzburg–Landau equation with fluctuating force reads [6].

$$\begin{aligned} &(\xi(x, t) \equiv \xi_u\,,\ \text{complex order parameter})\\ &\dot{\xi} = \lambda\xi + a\Delta\xi - c\,|\xi|^2\,\xi + F(t)\,. \end{aligned} \tag{16}$$

A further example is given by the Swift–Hohenberg equation [135], see also [24] (which was derived differently, however)

$$\dot{\xi}(x, t) = (a - b\Delta)^2\xi(x, t) + c\xi(x, t) - d\xi(x, t)^3\,. \tag{17}$$

The Eqs. (12,13,16,17) allow for a great variety of solutions. In the case of real λ and a single order parameter, a nonequilibrium phase transition occurs (see below). In case of λ complex, and (at least) one complex order parameter, Landau–Hopf bifurcation [67,68], i. e. formation of a limit cycle may happen. In case of (at least) three order parameters and no noise, deterministic chaos may occur [89,118,132] (in the presence of noise, mixed effects may occur).

Fokker–Planck Equation Below and above the instability point in control parameters space, in a first step the fluctuations can be neglected and then, in the next step, taken care of by means of lowest order perturbation theory. In order to cover the transition region, under well defined conditions a Fokker–Planck equation for the probability density function $f\left(\{\xi_u\}\right)$ of the order parameters can be derived. For details see [62,63a}] and the article by T. Frank, this volume.

The Fokker–Planck equation is of the general form

$$\dot{f}\left(\{\xi_u\}\right) = -\sum_u \frac{\partial}{\partial \xi_u}\left(\tilde{N}_u f\right) + \frac{1}{2}\sum_{uv}\frac{\partial^2}{\partial \xi_u \partial \xi_v}\left(Q_{uv} f\right)\,. \tag{18}$$

It is assumed that $\tilde{F}_u$ in (12) is δ correlated in time,

$$\langle \tilde{F}_u(t)\tilde{F}_v(t')\rangle = Q_{uv}\delta(t - t')\,. \tag{19}$$

If $\tilde{F}_u$ depends on ξ_u, the $\hat{I}$ to, Stratonovich or Klimontovich procedure must be applied.

In the case of a single order parameter, where the Langevin equation [88], originally with $\tilde{N} = -\alpha\xi$) is given by

$$\dot{\xi} = \tilde{N}(\xi) + F(t)\,, \quad \langle F(t)F(t')\rangle = Q\delta(t - t')\,. \tag{20}$$

The steady state distribution function of (18) is given by [62,116]

$$f(\xi) = N\exp\left(-2\int^{\xi}(\tilde{N}(\xi')/Q\mathrm{d}\xi') \equiv N\exp(-2V(\xi)/Q\right) \tag{21}$$

provided the boundary conditions are

$$f(\xi) \to 0 \quad \text{for } |\xi| \to \infty\,. \tag{22}$$

In the second Eq. (21), $Q = $ const. is assumed. N is a normalization constant. A generalization of (18) to continuous variables, $\xi_u(x, t)$, gives rise to a functional Fokker–Planck equation. An explicit solution of the Fokker–Planck equation in the case of several discrete or continuous order parameters can be found if the drift and diffusion coefficients obey the rules of detailed balance [45,47].

Nonequilibrium Phase Transition. Connection with Landau Theory The explicit form of the solution of the Fokker–Planck Eq. (21) allows us to make contact with the theory of phase transitions in the sense of the Landau theory [86] where

$$\begin{aligned} f(\xi) &= N\exp\left(-F(\xi, T)/(kT)\right)\,,\\ F(\xi, T) &= F(0, T) + a(T - T_c)\xi^2 + \frac{\beta}{4}\xi^4\,. \end{aligned} \tag{23}$$

In (21), V corresponds to the free energy F and the noise strength Q corresponds to *absolute* temperature T. T_c is the critical temperature, and (23) refers to a second order phase transition. In case of a first order phase transition, an additional term $\gamma\xi^3$ appears in (23).

An important difference between phase transitions at thermal equilibrium and in the present case of non-equilibrium should be mentioned, however. The decisive constants in the case of non-equilibrium [62] phase transitions are rate constants in contrast to thermodynamic quantities in (23). While non-equilibrium phase transitions described by (21) were experimentally very well verified for instance in the case of lasers [116] (Fig. 1), in the case of thermal equilibrium the Landau theory can not be considered as a good approximation and had been replaced by the concept of critical exponents etc. as dealt with by renormalization group theory [75,146]. For a treatment of the time dependent Fokker–Planck equations see Risken [117].

In a number of cases the drift- and diffusion coefficients of the Fokker–Planck equation are by themselves expectation values, defined on the probability density function so that the Fokker–Planck equation becomes non-linear. For more details see the article by T.D. Frank in this volume.

Instability of a Limit Cycle, $q_0(t)$ [62] The instability is checked by linear stability analysis by means of the hypothesis

$$q(t) = q_0(t) + W(t)\,, \tag{24}$$

where $q_0(t)$ is a time-periodic solution to (1) with $\alpha = \alpha_0$, $W(t)$ a small deviation.

Inserting (24) into (1) with $F \equiv 0$ and linearization leads to an equation of the form (4), where L because of $q_0(t)$ has become also a time-periodic function with the same period as $q_o(t)$. According to Floquet theory [35], the solutions to (4) with periodic $L(t)$ are given by

$$W(t) = \mathrm{e}^{\lambda_f t} v_j(t) \tag{25}$$

(in the case of nondegeneracy), where $v_j(t)$ has the same period as q_0, i. e. L.

Depending on $\operatorname{Re}\lambda_j \geq 0$ or < 0 we distinguish between unstable and stable modes (6, 7), respectively. One eigenvalue is $= 0$ and corresponds to an indeterminate phase shift, which in nonlinear analysis is taken care of by a phase $\phi(t)$ that acts as additional order parameter. In order to solve the fully nonlinear and stochastic equations, the hypothesis

$$q(t) = q_0(t + \phi(t)) + \sum_u \xi_u(t) v_u\left(t + \phi(t)\right) + \sum_s \xi_s(t) v_s\left(t + \phi(t)\right) \tag{26}$$

is inserted in the Eqs. (1). The subsequent procedure follows the lines outlined above and leads to order parameter equations of the form

$$\dot{\xi}_u = \lambda_u \xi_u + \hat{N}_u\left(\{\xi_u\}, \phi\right) + \hat{F}_u\left(\{\xi_u\}, \phi\right) \tag{27}$$

$$\dot{\phi} = M\left(\{\xi_u\}, \phi\right) + G\left(\{\xi_u\}, \phi\right) \tag{28}$$

where $\hat{N}, \hat{F}, M, G$ are polynominals in $\{\xi_u\}$ and periodic functions of ϕ.

The novelty as compared to the case of an unstable fixed point consists in the introduction of a phase as order parameter.

When noise is neglected, the newly evolving, i. e. bifurcating solutions are either two (or several) limit cycles or tori. Also basically, depending on the system, also a "back bifurcation" to a stable focus can happen.

Instability of Tori [62] The corresponding theory is rather complex so that a few words must suffice here. The basic idea [62] is based on an extension of (24,26) where q_0 is chosen as a quasi periodic function

$$q_0 = q_0(\omega_1 t, \omega_2 t, \ldots, \omega_M t) \tag{29}$$

where the $\omega' s$ must be sufficiently irrational in the sense of the KAM (Kolmogorov [80], Arnold [3], Moser [100]) theorem. Besides amplitudes as order parameters, also phases $\phi_1(t), \ldots, \phi_M(t)$ are introduced. For details cf. [62], and for alternative approaches [21,128].

A Remark on the Method of Solution of Evolution Eqs. (1)

In this article the central role of order parameters is stressed because this allows us to establish profound analogies between quite different systems. In practical applications it may be preferable, however, to apply other methods of solution, analytical, numerical or mixed, in order to derive the spatial, temporal or spatio-temporal patterns. In this way, the Springer Series in Synergetics have developed a "tool box" of models [98].

Quantum Theoretical Formulation

In a quantum theoretical treatment one deals with quantum mechanical Langevin equations which are Heisenberg equations of motion for operators to which pumping and damping terms as well as random noise sources are added. Here, according to quantum theory, the system's observables are represented by time-dependent quantum mechanical operators, Ω_j. For instance, by the position operator $\hat{x}$ and the momentum operator $\hat{p}$ of a particle, or, in quantum field theory, by creation and annihilation operators $\hat{b}^+, \hat{b}$, respectively. The quantum mechanical Langevin equations read (see, for instance [54,58]):

$$\dot{\Omega}_j = \frac{i}{\hbar}\left[H, \Omega_j\right] + \text{damping} + F_j(t)\,, \qquad (30)$$

where H is the Hamilton operator, and $F_j(t)$ are stochastic operators which usually are assumed to be δ-correlated in time. The quantum mechanical properties can be determined by the postulate of quantum mechanical consistency of Ω_j, (cf. [54], appendix).

If the non-commutativity of operators is taken care of, the procedure to derive order parameter equations is formally the same as in the case of classical Langevin equations as indicated above. The Fokker–Planck equation, however, must be replaced by a density matrix equation, originally introduced as master equation [109]. For nonequilibrium systems, such as the laser, see [50,126,144], also [54,119]. Using methods of quantum classical correspondence, this density matrix equation can be converted into a Fokker–Planck equation under specific conditions. The basic idea is this:

Quantum-Classical Correspondence

There are several ways to define quantum classical correspondence. In the case of position operator $\hat{x}$ and momentum operator $\hat{p}$ with the commutator $[\hat{p}, \hat{x}] = \frac{\hbar}{i}$ and the density matrix ρ, the Wigner distribution function $W(x, p)$ [145] is defined by

$$W(x, p) = \frac{1}{(2\pi)^2} \cdot \int \int_{-\infty}^{\infty} \mathrm{e}^{-ikx-ilp} \cdot \operatorname{tr}\left(\mathrm{e}^{ik\hat{x}+il\hat{p}}\rho\right) \mathrm{d}k\mathrm{d}l \qquad (31)$$

where "tr" means trace.

Thus a relation is established between the quantum mechanical density matrix and a classical quasi-density $W(x, p)$. Based on (31) or (34,35,36), a density matrix equation can be converted into a generalized Fokker–Planck equation [54].

By the transformation of $\hat{x}, \hat{p}$ to creation and annihilation operators b^+, b by means of

$$\hat{b}^+ = \frac{1}{\sqrt{2\hbar}}(\hat{x} + i\hat{p}) \qquad (32)$$

$$\hat{b} = \frac{1}{\sqrt{2\hbar}}(\hat{x} - i\hat{p}) \qquad (33)$$

an alternative form to (31) is given by

$$P(\beta, \beta^*) = \frac{1}{\pi^2} \int \int_{-\infty}^{\infty} \mathrm{e}^{-i\beta k - i\beta^* l} \cdot \operatorname{tr}(\mathrm{e}^{ik\hat{b}^+ + il\hat{b}}\rho)\mathrm{d}k\mathrm{d}l\,. \qquad (34)$$

Because $\hat{b}^+, \hat{b}$ are noncommuting operators, $[\hat{b}^+, \hat{b}] = 1$, different "quasiprobability" distributions P result, if

$$\mathrm{e}^{ik\hat{b}^+ + il\hat{b}}$$

is replaced by

$$\mathrm{e}^{ik\hat{b}^+}\mathrm{e}^{il\hat{b}} \qquad (35)$$

or

$$\mathrm{e}^{ik\hat{b}}\mathrm{e}^{il\hat{b}^+}\,. \qquad (36)$$

(35) gives rise to the Glauber–Sudarshan representation. For details and references see [54].

Regular Spatial and Spatio-Temporal Patterns

One of the most striking features of nonequilibrium systems in physics, chemistry and biology is their capability of forming (more or less) regular spatial pattern (for explicit examples see below). (There is a rich literature on pattern formation in physics, especially fluids [20,24,92,136], but also semiconductors [125] and nonlinear optics [133], chemistry [28,33,83] and biology [7,95,96,102] and general [69,70,98,104,110,111,115,140]. Furthermore, the patterns exhibit striking similarities in spite of the fact that the individual parts are quite different. The methodology of Synergetics (e. g. [62]) provides us with a basic insight into the causes of such analogies.

Pattern formation is determined by at least three causes:

1. internal mechanism, such as e. g. the interplay between reactions and diffusion in large scale chemical processes,
2. the influence of boundaries,
3. initial conditions.

Concerning 1) and 2) between two (limiting) cases can be distinguished.

1. dimensions of the internally evolving patterns are of the same or larger order as those of the boundaries. Here a strong influence of the boundaries must be expected.

2. dimensions of evolving patterns are small compared to those of the boundaries (boundaries → ∞).

To bring out the essential features we consider that originally for a control parameter value α_0 the system is homogeneous and quiescent. The approach can, however, be extended to a space dependent reference state (which, e. g. resulted from a first bifurcation leading to $q_0 = q_0(x)$) and the cases of a limit cycle or torus. The space may be 1, 2 or three dimensional Euclidian or, e. g., a 2 or 3 sphere.

Infinite Boundaries

We start with 2) infinite boundaries, the medium is homogeneous and isotropic. We assume a continuous transition from the homogeneous to the "bifurcating" state. The evolving patterns are determined by the leading terms in (8) that we call the "mode skeleton"

$$q(x,t) = q_0 + \sum \xi_u(t) v_u(x) \tag{37}$$

and the order parameter Eq. (12). The functions $v_u(x)$ are *the space-dependent part of the* solutions to (4) where L is a differential (or integral) operator which is invariant against translation and rotation. Thus, e. g., L commutes with the displacement operator

$$\Omega_a\colon\ x \to x + a\,, \qquad a \text{ constant vector}\,.$$

Thus $v_u(x)$ can be chosen as eigenfunction to Ω_a,

$$\Omega_a v_u(x) = \Lambda v_u(x) \tag{38}$$

with

$$v_u = \mathrm{e}^{ikx} \tag{39}$$

$$\Lambda = \mathrm{e}^{ika} \tag{40}$$

i. e. plane waves. Which waves must be considered in (37) is determined by λ_u in (6) as well as by the order parameter Eq. (12).

The condition $\operatorname{Re}\lambda_u(k) = 0$ defines $k = k_{\text{crit}}$. As was shown by means of many examples $k \approx k_{\text{crit}} \neq 0$. If the boundaries are finite, such a discrete k must be chosen which comes closest to k_{crit}. If the boundaries tend to infinity, a continuous set k is taken care of by (generalized) Ginzburg–Landau equation (see above). If the boundaries are "narrow" in 1 or 2 dimensions, but large in the remaining dimensions, the wave vector k must be split into k_{II} and $k_\perp$ where k_{II} is practically continuous and $k_\perp$ discrete. Quite often only one $k_\perp$ (the most critical) needs to be considered. This leads to practically 2 (or 1) dimensional patterns connected with k_{II}. In the 2-dimensional case, the modes with $|k_{\text{II}}| = k_{\text{crit.}}$ are degenerate. This degeneracy can be lifted by a weak influence of boundaries (leading to roll patterns), by specific initial condition which (by chance) prefers a specific roll pattern, or by terms in the order parameter-equations that lead to specific combinations, e. g.

$$k_1 + k_2 + k_3 = 0\,, \tag{41}$$

where $k_j, j = 1,2,3$ belong to k_{II}.

This gives rise to the formation of hexagons. This is the case if the leading term of $\tilde{N}$ contains

$$\int v_{k_1} v_{k_2} v_{k_3} \mathrm{d}^2 x \neq 0\,. \tag{42}$$

In three dimensions this mechanism may lead to plane wave fronts stabilizing each other which gives rise to icosaeders, as observed in diatomea.

An important class of spatio-temporal patterns (in 2 dimensions) results when the system utilizes *rotation symmetry.* This can best be explained by the following example:

In many cases of practical interest, N in (1) and thus L in (4) contain the Laplace operator Δ. When written in planar polar coordinates r, ϑ, solutions to (4) are of the general form

$$v \propto \mathrm{e}^{i(m\vartheta - kr - \omega t)} \tag{43}$$

(times a rotation symmetric function $g(r)$) which represents *spirals*. $m = 0$ represents concentric rings, while an integer $m > 0$ represents the number of spiral arms. $\omega = 0$ represents standing spirals, $\omega \neq 0$ rotating spirals.

The mode skeleton (37) is composed of functions of the form (43). Which of the functions (43) appear in (37) depends on the competition Eqs. (12) for order parameters, which may also allow for a super position of counter rotating spirals (such as in the sunflower head). As group theory shows (see below), solutions (43) with different m's belong to different irreducible representations, and do not coexist in (37). This does not exclude the coexistence of differently rotating spirals in *different* regions of space, however.

The above results can be cast into the isomorphy principle:

While the "true" q is represented by (we omit the homogeneous q_0)

$$q = \sum_u \xi_u v_u(x) + \text{enslaved modes, with same symmetry.} \tag{44}$$

and v_u "true modes", its symmetry features can be replaced by a "representative" q':

$$q' = \sum_k \xi_k R_k(x)\,, \tag{45}$$

where R_k represent the "elementary" functions showing the symmetry under consideration. While the material significance and explicit form of q according to (44) may be quite different for different material substrates, q' (45) shows the *same* patterns for different systems.

These results can be deepened by invoking group theory, in which also the effect of the boundaries is taken into account.

Symmetries, Group Theory, Representation Theory, Finite Boundaries

Consider a set of transformations G_j of space variables $x \to x'$ so that

$$G_j q \to q' \tag{46}$$

Example 1 G_j induces the translation

$$x \to x + a \quad \text{so that} \quad G_j q(x) = q(x + a)\,. \tag{47}$$

The transformations must be so that they are compatible with the internal properties of the system (1) and the boundary conditions. Example: when dealing with a problem on a 2-dimensional sphere, the transformed coordinates x must not leave the sphere.

Because of the symmetry of the problem, the transformations G_j form a group defined by

1. existence of unity E such

$$G_j E = G_j \quad \text{for all } j \tag{48}$$

2. the product of two group elements is again an element of the group,

$$G_j G_k = G_l \quad \text{for all } j, k \tag{49}$$

3. existence of an inverse G_j^{-1} for all j so that

$$G_j^{-1} G_j = E\,, \tag{50}$$

4. associative law

$$(G_k G_l) G_j = G_k (G_l G_j) \tag{51}$$

for all group elements.

In the following we first ignore random forces, i. e. we consider (1) with $F \equiv 0$.

$$\dot{q}(x, t) = N(q, \Delta, \alpha)\,. \tag{52}$$

Jointly with the boundary conditions, (52) defines a function space S in which all functions to be considered must lie (i. e. can be represented by linear combinations of a complete set of (vector valued) basic functions of S; example: S is a Hilbert space)

Definition 1 The system is invariant against G_j if for all $f \varepsilon S$

$$G_j N\left(G_j^{-1} f\right) = N(f)\,. \tag{53}$$

Example 2

$$G_j\colon\ x \to x + a\,, \tag{54}$$

$$N(f) = \Delta f + V(x) f + f^2\,. \tag{55}$$

Then

$$G_j{\cdot}N\left(G_j^{-1} f\right) = \Delta G_j^{-1} f(x{+}a) {+} V(x{+}a) G_j^{-1} f(x{+}a) + \left(G_j^{-1} f(x + a)\right)^2 \tag{56}$$

$$= \Delta f(x) + V(x + a) f(x) + f(x)^2 \tag{57}$$

$$\neq N(f) = \Delta f(x) + V(x) f(x) + f(x)^2 \tag{58}$$

unless $V(x + a) = V(x)$. If a in (54) is arbitrary, N is not invariant against (54).

Application of G_j to q in (52) leads to

$$\frac{\mathrm{d}}{\mathrm{d}t}(G_j q) = N(G_j q) \tag{59}$$

or because of (53), (with $f = G_j q$), to

$$\frac{\mathrm{d}}{\mathrm{d}t}(G_j q) = G_j N(q)\,. \tag{60}$$

In the spirit of representation theory of groups the action of G_j on f can be understood as an abstract operation, but also as a matrix acting on the vector f in S- space.

By appropriate transformation of basis of q, and using the representation theory of symmetry groups, all matrices U_j belonging to all group elements j can simultaneously be decomposed into "irreducible" representations so that (in the example of 3 irreducible representations)

$$U_j = \begin{pmatrix} \square & \bigcirc & \bigcirc \\ \bigcirc & \square & \bigcirc \\ \bigcirc & \bigcirc & \square \end{pmatrix} \tag{61}$$

Each box □ is a matrix $U_j^{(k)}$ with dimension Dk, so that

$$D1 + D2 + \cdots + Dk = \text{dimension } U_j \,.$$

Example 3 Rotation group applied to 2-sphere (e. g. earth surface). Basis functions are spherical harmonics Y_m^l with "quantum numbers" l, m. Subspace l fixed, $m = 0, \ldots, l-1$. As a consequence, the mode skeleton reduces to (q_0 dropped)

$$q^l = \sum_m \xi_m(t) Y_m^l \,. \tag{62}$$

There is no coupling between different ls, which implies a low dimensional dynamics of ξ_m.

Generally, the original function space S is decomposed into subspaces forming the basis of each irreducible representation. This implies a symmetry reduction beyond bifurcation point, compared to the situation below bifurcation point, where

$$G_j q = q \quad \text{for all } j \,, \tag{63}$$

i. e. q fully symmetric under G.

In our example beyond the bifurcation point q is given by q^l where Y_m^l transforms according to the subgroup G^l, which leaves the space spanned by Y_m^l invariant. If, however, group elements not belonging to G^l are applied to q^l, this space is left. In other words, q^l is connected with a lower symmetry than q (63). By bifurcations, the symmetry of q is lowered and one speaks of "symmetry breaking instability". If fluctuating forces in (1), i. e. in (52) are taken into account, the full symmetry can be restored (under specific conditions on the fluctuating forces).

While group theory has found important and widespread applications to quantum theory, it is less frequently used in problems of Synergetics, though there it may lead to deep insights as pointed out above. (For an in-depth approach see [42,43,120].)

On top of, or jointly with, regular patterns, a variety of defects as well as boundaries between different patterns may occur (cf. contribution by Pismen, this volume and [110,111]).

A Further Mathematical Tool: Shannon Information and the Maximum (Information) Entropy Principle

While evolution equations are the backbone of Synergetics, also other tools are invoked to deal with complex systems. Such a tool is Shannon information [129] which is defined by

$$i = -\sum_j p_j \log_2 p_j \tag{64}$$

where p_j is the relative frequency of the event j or, in a different interpretation, the probability of finding the realization j in an experiment. The maximum (information) entropy principle as formulated by Jaynes [73,74], for an earlier proposal see [27]), allows one to make unbiased guesses on systems on which only incomplete data are known by maximizing the informations, i. e. (64) = max! or = extremum! under given constraints.

A simple example is provided by a gas composed of N particles, where the total kinetic energy $\mathrm{E}_{\text{kin}}^{\text{tot}}$ is fixed. Denoting the kinetic energy of a particle with mass m and velocity v_i by $f_i = (m/2)v_i^2$, the mean kinetic energy per particle is

$$\sum_i p_i f_i = \mathrm{E}_{\text{kin}}^{\text{tot}}/N \tag{65}$$

To fix p_i, (64) must be maximized under the normalization condition

$$\sum_i p_i = 1 \tag{66}$$

and the constraint (65).

Using Lagrange multipliers, λ, λ_1, the result reads

$$p_i = \exp\left(-\lambda - \lambda_1 m v_i^2/2\right) \tag{67}$$

i. e. the Maxwell–Boltzmann distribution function. Also relations between the Lagrange multipliers λ, λ_1 can be established which, evidently, have fundamental thermodynamic significance.

This approach has been extended to the treatment of nonequilibrium phase transitions, i. e. determination of order parameters, enslaved modes and emerging patterns [60]. The crucial idea consists in the proper choice of constraints, as which the moments of the variables q_i are chosen:

$< \ldots >$ means average over the joint distribution function $f(q_1, q_2, \ldots, q_n)$ which replaces p_j and the vector $(q_1, \ldots, q_N)$ replaces j. The variables q_j may be discrete or continuous.

$$f_i =< q_i >, \quad i = 1, 2, \ldots, N \,. \tag{68}$$

$$f_{ij} =< q_i q_j > \,. \tag{69}$$

$$f_{ijkl} =< q_i q_j q_k q_l >, \quad i, j, k, l = 1, 2, \ldots, N \,. \tag{70}$$

The resulting distribution function is given by

$$q = \exp V(\lambda, q) \tag{71}$$

with

$$V(\lambda, q) = \lambda + \sum_i \lambda_i q_i + \cdots + \sum_{ijkl} \lambda_{ijkl} q_i q_j q_k q_l \,. \quad (72)$$

(71) is a starting point to make contact with the Landau or Ginzburg–Landau theory of phase transitions [86], and to guessing Fokker–Planck equations. The approach allows one to calculate the efficiency of self-organizing systems close to their instability points.

The method has been extended to the "unbiased modeling" of stochastic processes: how to guess path integrals, Fokker–Planck equations and Langevin-Îto equations [60]. The central quantity to be searched for is the probability density P_n of paths.

Let $q(t)$ be the state vector $q = (q_1, \ldots, q_N)$ at time t, then

$$P_n(t_n, t_{n-1}, \ldots, t_0) = P_n\left(q(t_n), t_n; q(t_{n-1}), t_{n-1}; \ldots ; q(t_0), t_0\right) , \quad t_n > t_{n-1} > \ldots > t_0 \,. \quad (73)$$

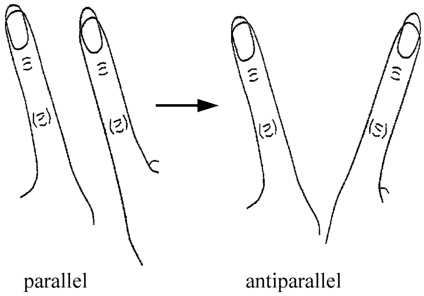
This task is simplified if the Markov hypothesis on the process holds, i. e.

$$P_n(t_n, t_{n-1}, \ldots, t_0) = \hat{P}\left(q(t_n), t_n \middle| q(t_{n-1}) t_{n-1}\right) \cdot P_{n-1} \quad (74)$$

where $\hat{P}$ is the transition probability so that only transition probabilities between subsequent states (with $\Delta t \to 0$) must be guessed in addition to P_0. In the frame of the present approach, this task is fulfilled by use of the maximum information principle. The constraints to be used are essentially conditional first order moments and two-time correlation functions of the state vectors $q(t)$, $q(t')$.

Phenomenological Synergetics

In many fields of science, including medicine, the microscopic variables and their dynamics are not well-known or not known at all. Nevertheless, in quite a number of cases, namely where dramatic macroscopic changes of the system's behavior take place, general insights, gained by Synergetics, can be invoked. A paradigm for this procedure is the modeling of Kelso's finger experiments [77,78] (Fig. 3). He instructed subjects to move their index fingers in parallel which was accordingly performed. However, when the speed of the fingers was increased, the parallel movement was replaced by a symmetric movement quite involuntarily and spontaneously. In other words, a transition from a parallel to an anti-parallel phase takes place. In terms of Synergetics, the interpretation is simple: the control parameter consists in the prescribed frequency ω of the finger movement, whereas the macroscopic quantity, i. e. the order parameter that changes dramatically is provided by the relative phase of the two index fingers. According to the experience made in Synergetics, the order parameter, here called ϕ obeys a typical order parameter equations of the form [64]



Synergetics: Basic Concepts, Figure 3
Transition between finger movements from parallel to symmetric in Kelso's experiment [64]

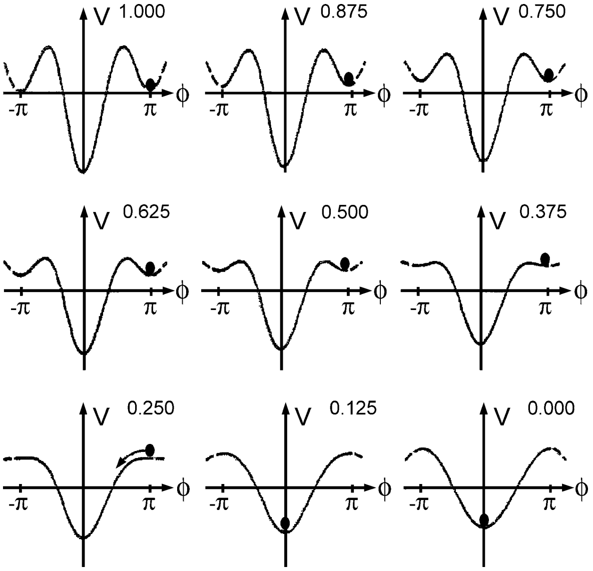
$$\dot{\phi} = -\frac{\partial V}{\partial \phi} + F(t) \,, \quad (75)$$

where $V(\phi, \omega)$is a potential function and F a fluctuating force. When the control parameter ω is changed, the potential runs through a series of forms as depicted in Fig. 4. As was shown in detail, at a critical value of ω, the transition from one potential minimum to another one occurs, as related to the change of the kind of finger movement. The mathematical analysis shows hysteresis, critical slowing down and critical fluctuations [59] which reject the idea that the brain acts like a computer via a motor program but rather via self-organization.

Another application is made by the Synergetic computer [61] (Figs. 5, 6), where to each pattern to be recognized a specific order parameter is attached. Pattern recognition is then achieved via a competition between order parameters. The competition equations are given by

$$\dot{\xi}_k = \frac{\partial V}{\partial \xi_k} V(\xi_1, \ldots, \xi_M) = -\frac{1}{2} \sum_k \lambda_k \xi_k^2 + \beta \sum_{k,k'} \xi_k^2 \xi_{k'}^2 - C \sum_k \xi_k^4 \,. \quad (76)$$

This approach may serve also for modeling of brain functions: both recognition as well as movements are governed by the establishing of order parameters which may wander from one quasi attractor to another one. Quasi attractors are defined as attractors that vanish after the task has been accomplished, e. g. after a pattern has been recognized or movement performed.



Synergetics: Basic Concepts, Figure 4
Sequence of potential curves of the Haken–Kelso–Bunz model of Kelso's experiment [64]

Based on the concept of order parameters, a learning procedure for Synergetic computers has been developed [61]. Here the number of patterns to be recognized is prescribed and then a special functional must be minimized. In the case of the Synergetic computer, it is possible to make contact between the microscopic and the mesoscopic description, i. e. the microscopic variables are pixel values q_j, j pixel index, whereas the mesoscopic (or macroscopic) quantities are the order parameters ξ_k.

The relation between ξ_k, q_j is given by

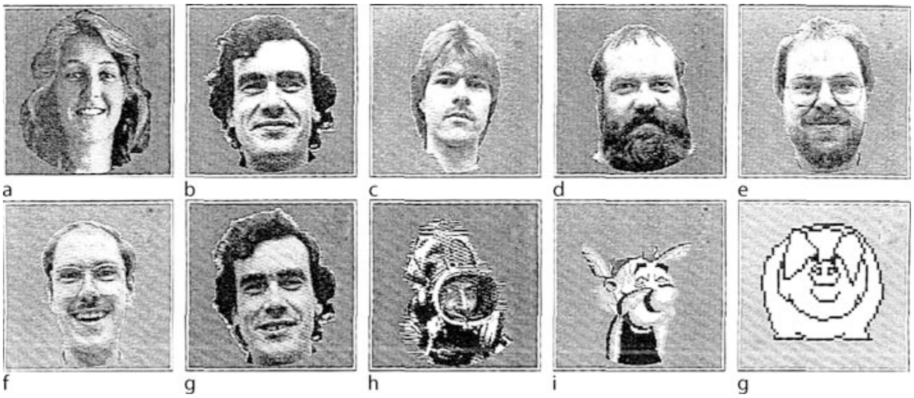
$$\xi_k = \sum_j v_j^{k+} q_j \,, \tag{77}$$

where v_j^{k+} are adjoint prototype patterns, with k pattern index, j pixel index.

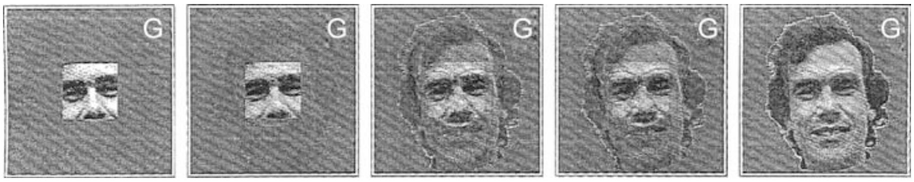
The relation between prototype patterns v_j^k and their adjoints is given by

$$\sum_j v_j^{k+} v_j^{k'} = \delta_{kk'} \tag{78}$$

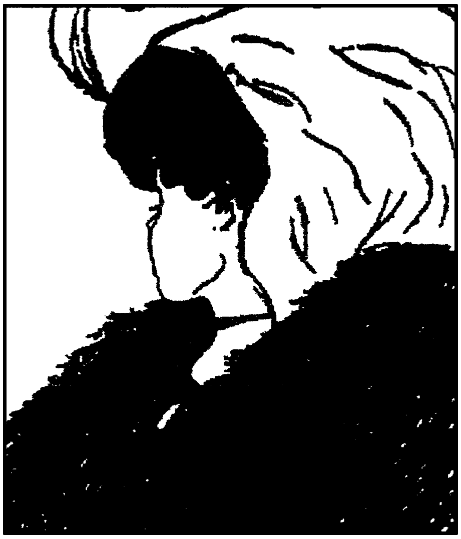
At the phenomenological level the order parameter concept allows us to interpret and model complex movement patterns, e. g. learning to ride on a pedalo [59]. In the experiments, LED's are fixed at the joints of the subject and their positions measured which gives rise to a series of time-dependent tracks. Then, in a first step, a principle component analysis is performed, in the next step, by means of a variational principle, the best fit is searched in



Synergetics: Basic Concepts, Figure 5
Recognition of faces by the synergetic computer: stored or learned prototype patterns [61]



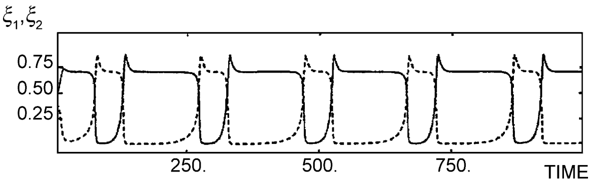
Synergetics: Basic Concepts, Figure 6
Pattern recognition by the synergetic computer: recognition of a specific face of which initially only a subset of pixels is presented [61]



Synergetics: Basic Concepts, Figure 7
Example of an ambivalent figure: young / or old woman? [34]

terms of order parameters and their equations of motion, in order to mimic the actual tracks. While in the learning phase several order parameters are needed, at the end the whole movement is governed by a rather simple equation for a complex order parameter.

During the development of Synergetics it turned out that there are strong relations to gestalt theory [85] as well as to psycho physics. A typical example is provided by ambivalent figures where (Fig. 7) [34] shows an example. An observer may either perceive a young woman or an old woman, but not both simultaneously, rather the perception switches between these two percepts. In the mathematical modeling to each percept an order parameter is attached [61], which obeys the typical equations of Synergetics. The control parameter invoked here is attention. According to an early suggestion by Wolfgang Köhler [85], when a pattern is recognized, the corresponding attention fades away. This has been modeled mathematically based on a competition dynamics between two order parameters, when the control parameter (attention) of one pattern fades away, the other pattern gets the possibility of being perceived. Then in the next step the corresponding attention parameter fades away and the first pattern may re-appear (Fig. 8) [61]. This model describes details of the observed phenomena, such as the dependence of the duration of the perception of one face as compared to that of the other face, dependent on the bias which face is recognized first. Also, one may distinguish between slow, medium and fast observers, depending on the individual parameters.



Synergetics: Basic Concepts, Figure 8
Order parameter oscillations belonging to the recognition young woman / old woman with bias towards the young woman [59]

Quite generally, order parameters may have properties of gestalt in the sense that they are invariant against size, orientation and perception of objects in space.

In medicine, a syndrome has the characteristic features of an order parameter. On the one hand it is generated by the co-operation, or at least by the simultaneous presence of specific features, on the other hand once the syndrome (order parameter) is established, it acts on the individual parts of the system, where the slaving principle induces specific phenomena at the level of individual parts. Clearly, the concept of circular causality plays an important role here. It shows that the syndrome, at least in general, can not be cured by curing an individual symptom, but rather by curing a decisive majority of individual causes.

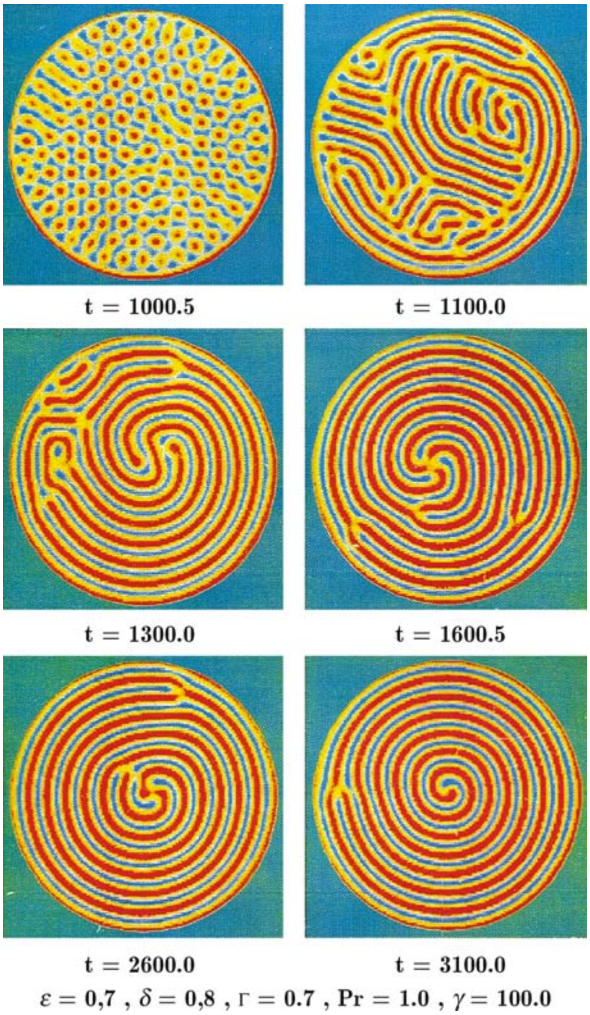
Semantic Synergetics

In soft sciences, but also in medicine and other fields, a mathematical modeling, even at the level of order parameters may not be possible. Nevertheless, Synergetics may provide us with qualitative insights into basic mechanisms. In psychology and psychiatry [122], quite often specific mental states can be ascribed to a patient. For instance in bipolar patients a depressive phase or a manic phase may appear or in depressive patients a normal phase and a depressive phase. Another example is provided by patients with a compulsory action. In the spirit of Synergetics, as a theory of indirect control, one may ask, whether there are appropriate control parameters by means of which the behavior of a person can be changed. Let the two states be represented by the positions of a ball in a landscape with two valleys. In this situation, direct control means to push the ball from the unwanted position to the wanted. Indirect control means to lower the potential hill between the two valleys so that the wanted transition may occur via self-organization. This may happen through interventions used in cognitive psychology, a change of environmental conditions, or/and by specific medication. The central issue here is that the patient is not directly influenced, e. g. by saying you must do this or that, but rather by a soft changing of his/her point of view. A number of successes have been reported about this method which is, to some extent, well known in psychiatry, but finds here a scientific theoretical basis. For more details see the article by G. Schiepek and V. Perlitz, this section, and in a somewhat related form [66].

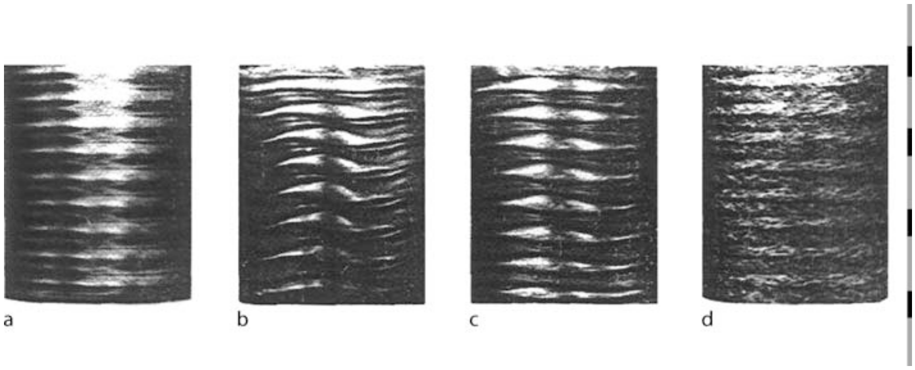
Some Selected Examples

The study of nonlinear, self-sustained oscillations [1,2,13] be it in radio-engineering, mechanics or other fields, has a long tradition. In the context of bifurcation theory, their origin was unearthed by Hopf [67,68].

Nonlinear optics [99] and, when quantum effects are important, quantum optics [57,97,123,142] provide us with a wealth of phenomena, in particular of the formation of coherent oscillations. A device, closely related to the laser, is the parametric oscillator [44], in which, within a nonlinear crystal, incoming pumplight is split into a signal and an idler. Then, similar to the laser light, the signal light becomes amplified, and its generation can be described as that of a nonlinear quantum-mechanical oscillator. Fluid dynamics is rich of pattern formations (including chaos) [12,16,17,18,31,41,89,92,105,115, 118,127,135,136], to mention just a few. In a fluid heated uniformly from below, with increasing temperature difference, several instabilities may occur for instance giving rise to stationary patterns, such as rolls, hexagons (Fig. 9) or squares. In the next step the rolls may start to show oscillations, and still more complex patterns may occur (Fig. 3). In the case of the Taylor instability [137], a liquid is placed in between two coaxial cylinders, where the outer one is rotating. With increasing rotation speed, a hierarchy of instabilities is reached, first the formation of roles, then oscillating rolls at one frequency, then oscillation of rolls at two frequencies, and finally weak turbulence, i. e. chaos occurs (Fig. 10) [31,93]. Important phenomena are the establishing of boundaries and of defects as described in the article by Pismen [110,111] and other articles of this Encyclopedia. A rich variety of pattern formation may occur in semiconductors [125], where electrons and holes as well as currents form specific spatio-temporal patterns. In meteorology, atmospheric convection patterns and other instabilities are treated [38]. In chemistry, oscillations and large scale patterns arise by means of the interplay of chemical reactions and diffusions [9,15,28,32,33,149], e. g. concentric ring patterns, each starting from a center, which then annihilate each other when colliding. An important class is provided by spiral patterns which may have one to several arms (Fig. 11). In biology, specific models on morphogenesis were treated, such as the formation of stripe or spot patterns on animal furs or skins of fish (Fig. 12) or still more complicated patterns on sea shells [39,62,95,96,102]. The basic idea which can be traced back to Turing [139] is this: originally unspecialized cells produce activator and inhibitor molecules which by reaction and diffusion form a prepattern, a morphogenetic field [147]. At positions of high activator concentration, genes are switched on which then leads to cell differention producing e. g. pigments. In aggregating slime mold, spiral or concentric ring patterns are observed [14,37]. Mathematical models on prebiotic evolution [26] study the competition between species of



Synergetics: Basic Concepts, Figure 9
Model calculation of the motion of a fluid in a circular pan uniformly heated from below (after [29]). *Upper left corner*: above a critical temperature difference between lower and upper surface of the fluid layer, a hexagonal pattern appears. If the boundary is also heated uniformly, a transition to the spiral pattern with one or several arms can be found (*lower right corner*)



Synergetics: Basic Concepts, Figure 10
Pattern hierarchy in the Taylor–Couette instability. A fluid in between two vertical coaxial cylinders of which the outer one rotates, shows no macroscopic movement pattern, if the movement of the outer cylinder is slow. When the rotation speed is increased, first a role pattern appears in which the fluid moves outwards at one height, and then inwards at another height. This movement pattern is periodic with respect to height [137]. At a further critical rotation speed, the pattern shows oscillations which at a further speed transform into a motion with two frequencies until eventually chaotic motion appears. The experiments were done by [31], the modeling was done for the first transition (homogeneous to roles) and especially the second transition (roles to oscillating roles) by [93]

Synergetics: Basic Concepts, Figure 11
Belousov–Shabotinsky reaction: the occurrence of spirals (courtesy A.T. Winfree). They may show one to several arms. The centers of the spirals may occur at different positions. Spirals hitting each other, annihilate each other

Synergetics: Basic Concepts, Figure 12
Stripe pattern on a tropical fish

biomolecules and the "survival" of the fittest, where pronounced analogies with the dynamics of laser photons can be unearthed, fully in line with Synergetics [62]. In the understanding of brain function, for instance, steering of movements, pattern recognition or decision making, the reduction of degrees of freedom of the numerous neurons to few order parameters is central [59].

The concepts and principles of Synergetics shed new light on important relationships in economy, such as cooperation and competition between companies, the important role of indirect steering by means of control parameters, such as taxes, interest rates. It can be shown, that a fusion of companies does not necessarily lead to so called synergy effects, but rather critically depends on initial conditions and details of the cooperation between the previously separated firms. Important insights are also gained into fundamental processes of climatology, as well as in ecology such as the by now well-known and publicly discussed effects that even small concentrations of chemicals in the atmosphere can change the climate dramatically. The same is true for lakes, in which beyond a critical pollution, fish population dies out entirely.

In this way, the numerous examples collected in the field of Synergetics, provide not only scientists but also the public with impressive examples of dramatic changes (instabilities) provoked by even a slight change of control parameters. Clearly, an important research subject of Synergetics is a detailed study of which control parameters are critical and to which control parameters a system is rather insensitive. Sociology is an important field for the application of stochastic models [8]. In particular, basic concepts of Synergetics have proven useful in the developing field of sociodynamics, where e. g. phase transition-like phenomena may occur [143].

History and Relations to Other Fields

The term Synergetics was coined by H. Haken in 1969 in a lecture at University of Stuttgart. A first description of the goals of this field was given by H. Haken and R. Graham in 1971 [63] where the unifying role of the concept of order parameters is outlined. A relationship exists to the general system theory due to von Bertalanffi [10], which also aims at the exploration of analogies between different systems, but on the level of the individual elements rather than on the level of order parameters. Von Bertalanffi coined the term flux equilibrium (Fließgleichgewicht) in order to characterize homeostasis in active systems [11]. A general mathematical frame for Synergetics is provided by dynamic systems theory (see, for instance, [49]) which, however, in the traditional approach ignores stochastic processes (mainly chance events) which are also of great relevance for Synergetics. Here the theory of Markov processes with their typical equations, such as Langevin equations, Fokker–Planck equations, Chapman–Kolmogorov equations, the Kramers–Moyal expansion etc. is important (see for instance [134] and ▶ Linear and Non-linear Fokker–Planck Equations by T. Frank).

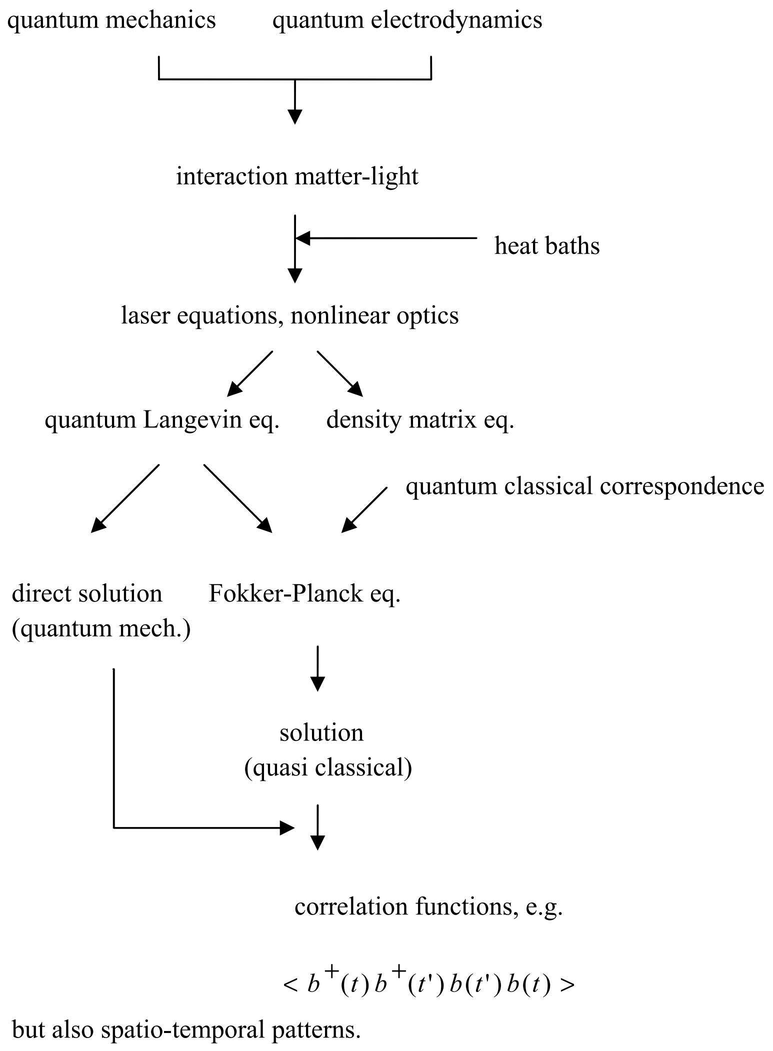
A basic feature of Synergetics consists in dealing with nonlinearities in complex systems and studying, mainly quantitatively, qualitative changes at macroscopic scales. Qualitative changes of systems at macroscopic levels are studied also by catastrophe theory [5,138], which may be interpreted as a study of the surfaces of equilibrium points of few order parameters, where different cases are classified according to the (low) number of control and order parameters. Chaos theory studies the mostly irregular dynamics of deterministic low dimensional continuous [89,106,118,132] or discrete dynamic systems [23,30,48,94,130], where the behavior is mainly characterized by so called Lyapounov exponents, various kinds of fractal dimensions and chaotic attractors. The slaving principle of Synergetics provides a basis for an application of chaos theory to multi-component systems in that Synergetics shows the possibility of reducing the degrees of freedom. Synergetics shares some of its topics with singularity theory [4,42,43], which applies to bifurcation points and their surrounding. Another point of contact is bifurcation theory (see the quotations in previous chapters), in which the branching of solutions of the dynamic system close to instability points is studied. The term dissipative structure was coined by Prigogine [40] to characterize evolving structures in systems away from thermal equilibrium where as in all such non-equilibrium systems dissipation occurs. A typical example is that of the convection instability. Prigogine tried to base his approach on thermodynamics, introducing concepts of entropy production and excess entropy production. As we now know, these concepts are, however, insufficient to deal with structure formation in such systems [87]. Based on a fundamental idea of A. Turing [139], Prigogine and Nicolis [114], see also [108], treated macroscopic pattern formation in a specific chemical reaction model. For more recent work see [107].

Because of the fundamental importance of thermodynamics, we elucidate its relationship to Synergetics more closely.

Thermodynamics (see for instance [19]) deals with systems in and out of thermal equilibrium. A central concept is entropy. In a closed system, it tends to its maximum value. Thermal equilibrium is characterized by the equipartition theorem: each degree of freedom has an average energy of $1/2\ kT$, $k =$ Boltzmann constant, T absolute temperature. This may refer e. g. to gas atoms as well as to collective excitations in crystals. These systems are in thermal equilibrium with their surrounding (heatbaths, reservoirs).

Irreversible thermodynamics [51] treats systems which are not in thermal equilibrium but close to it. It mainly deals with transport and relaxation processes. A central concept is entropy production.

In the domains of physics, chemistry, biology, Synergetics deals with systems far from (thermal) equilibrium . This state is caused and maintained by an in- and outflow of matter, energy and / or information. This is achieved by a coupling of the “proper system” to heat baths (reservoirs) at different temperatures. The former concepts of thermodynamics, in particular the first and second law, are still valid for the total system (“proper” plus reservoirs), but no more sufficient to deal with the kinetics of the proper system. Now the central concept is growth and decay rates. In systems far from thermal equilibrium, collective modes are formed. One or several of them compete best for the external supply of matter, energy, information and grow at the expense of all other degrees of freedom (or modes). Thus the equipartion theorem is no more valid. In general, the behavior of the system is governed by few degrees of freedom (order parameters). Inci-


quantum mechanics
quantum electrodynamics
interaction matter-light
heat baths
laser equations, nonlinear optics
quantum Langevin eq.
density matrix eq.
quantum classical correspondence
direct solution (quantum mech.)
Fokker-Planck eq.
solution (quasi classical)
correlation functions, e.g.
$< b^{+}(t) b^{+}(t') b(t') b(t) >$
but also spatio-temporal patterns.

Synergetics: Basic Concepts, Scheme 1
Quantum optics, example laser (after [54])

dentally, this "growth and competition" principle applies to a great variety of fields out of physics, chemistry and biology, where "modes" may not only be special physical structures, but may mean behavioral patterns, special functions etc. Quite often, a "mode" is initiated by a chance event (fluctuation). Clearly, a generalized Darwinian principle can be seen: The interplay between *mutations* (microscopic chance events) and *selection* (competition between mascropic modes) leads to macroscopic patterns (structures) in the widest sense of the word.

In present days research, a new name is spreading, namely complexity or complexity theory. There seems to be no precise definition of this field available in the scientific community. Of what is known so far, we may conclude that this field has strong ties to the original field of Synergetics in that it searches also for general principles but, in addition, it allows the collection or accumulation of knowledge on all kinds of complex systems, as is witnessed in the excellent Complexity Digest, weekly edited by Gottfried Mayer. What "complexity" eventually might mean is reflected by the present encyclopedia.

Future Directions

Synergetics is surely not a closed scientific discipline but quite open to further research. On the one hand we may think of further applications of the principles of Synergetics that have been hitherto elaborated on, such as order parameters etc. Here a wide field of application is provided by robotics, construction of prostheses, automatic steering of cars etc. On the other hand, new ideas to endow systems with self-organizing properties are needed, e. g. groups of mobile agents for the execution of specific tasks. First steps have been done for instance by Kornienko [81].

Bibliography

Basically, the individual volumes of the Springer Series in Synergetics and selected volumes of the Springer Series on Complexity provide further information. Here, because of lack of space, only a few basic papers can be quoted.

The number of books, reviews and original papers on the topic of the present article is enormous. I have tried to achieve a fair balance between original contributions and reviews / books. Nevertheless, my selection must remain – to some extent – arbitrary and incomplete.

Primary Literature

1. Abraham R, Marsden JE (1978) Foundations of mechanics. Benjamin / Cummings, Reading
2. Andronov A, Vitt A, Khaikin SE (1966) Theory of oscillators. Pergamon Press, London-Paris
3. Arnold VI (1963) Russ Math Surv 18:9
4. Arnold VI (1993) Dynamical systems VI singularity theory 1 v 6 (Encyclopaedia of Mathematical Sciences). Springer, Berlin
5. Arnold VI, Afrajmovich VS, Il'yashenko, Yu S, Shil'nikov Lf (1999) Bifurcation theory and catastrophe theory. Springer, Berlin
6. Aronson IS, Kramer L (2002) The world of the complex Ginzburg–Landau equation. Rev Mod Phys 74:99–143
7. Babloyantz A (1986) Molecules, dynamics and life: An introduction to Self-Organization of matter. Wiley, Indianapolis
8. Bartholomew DJ (1967) Stochastic models for social processes. Wiley, London
9. Belousov BP (1959) Sb Ref radats Med, Moscow
10. Bertalanffi L von (1950) Brit J Phil Sci 1:134
11. Bertalanffi von L (1953) Biophysik des Fließgleichgewichts. Vieweg, Braunschweig
12. Bodenschatz E, Pesch W, Ahlers G (2000) Annu Rev Fluid Mech 32:709
13. Bogoliubov NN, Mitropolsky YA (1961) Asymptotic methods in the theory of nonlinear oscillations. Hindustan Publ Corp, Delhi
14. Bonner JT, Barkley PS, Hall EM, Konjin TM, Mason JW, O'Keefe G, Wolfe PB (1972) Devel Biol 20:72
15. Bray CH (1921) J Am Chem Soc 43:1262
16. Busse FH (1972) J Fluid Mech 52:1–97
17. Bénard H (1900) Ann Chim Phys 7(23):62
18. Bénard H (1900) Rev Gén Sci Pures Appl 11:1261–1309
19. Callen HB (1960) Thermodynamics. Wiley, New York
20. Chandrasekhar S (1961) Hydrodynamic and hydromagnetic stability. Clarendon Press, Oxford
21. Chenciner A, Iooss G (1979) Arch Ration Mech Anal 69:109
22. Chow S-N, Hale JK (1982) Methods of bifurcation theory. Springer, Berlin
23. Collet P, Eckmann JP (1980) Iterated maps on the interval as dynamical system. Birkhäuser, Boston
24. Cross MC, Hohenberg P (1993) Rev Mod Phys 65:851
25. DeGiorgio V, Scully MO (1970) Phys Rev A 2:1170
26. Eigen M, Schuster P (1977) Naturwissenschaften 64:541, (1978) 65:7, 65:341
27. Elsasser W (1937) Phys Rev 52:987, (1968) Z Phys 171
28. Epstein IR, Pojman JA (1998) Introduction to nonlinear chemical dynamics. Oxford University Press, New York
29. Fantz M, Bestehorn M Friedrich R, Haken H (1993) Phys Lett A 174:48–52
30. Feigenbaum MJ (1978) J Stat Phys 19:25
31. Fenstermacher RP, Swinney HL, Gollub JP (1979) J Fluid Mech 94:103
32. Field RJ, Korös E, Noyes RM (1972) Am Chem Soc 49:8649
33. Fife PC (1979) Mathematical aspects of reacting and diffusing systems. Springer, Berlin
34. Fisher G (1967) Measuring ambiguity. Am J Psychol 80:541–547
35. Floquet G (1883) Sur les équations différentielles linéaires à coefficients périodiques. Ann Ècole Norm 2(12):47
36. Gardiner CW (1994) Handbook of stochastic methods. Springer Series in Synergetics, vol 13. Springer, Berlin
37. Gerisch G, Hess B (1974) Proc Nat Acad Sci (Wash) 71:2118
38. Giaiotti DB, Steinacker R, Stel F (2007) Atmospheric convection. Research and Operational Forecasting Aspects, Cism International Centre for Mechanical Sciences Courses and Lectures. Springer, Wien
39. Gierer A, Meinhard H (1972) Kybernetik 12:30
40. Glansdorff P, Prigogine I (1971) Thermodynamic theory of structure, stability, and fluctuations. Willey, New York
41. Gollup J, Benson SV (1979) In: Haken H (ed) Pattern Formation by Dynamic Systems and Pattern Recognition. Springer Series in Synergetics, vol 5. Springer, Berlin
42. Golubitsky M, Schaeffer D (1988) Singularities and groups in bifurcation theory I, vol 1. Springer, Berlin
43. Golubitsky M, Stewart I, Schaeffer D (1988) Singularities and groups in bifurcation theory, vol 2. Springer, Berlin
44. Graham R (1970) Quantum statistics of optical parametric oscillation. In: Kay SM, Maitland A (eds) Quantum Optics. Academic Press, New York
45. Graham R (1981) Z Phys B 40:149
46. Graham R, Haken H (1968) Z Phys 213:420 (1970) 235, 237:31,166
47. Graham R, Haken H (1971) Z Phys 248:289
48. Grossmann S, Thomae S (1977) Z Naturforsch 32A:1353
49. Guckenheimer J, Holmes P (1983) Nonlinear oscillations, dynamical systems, and bifurcations of vector fields. Springer, Berlin
50. Haake F (1973) In: Springer Tracts in Modern Physics, vol 66. Springer, Berlin, p 98
51. Haase R (1969) Thermodynamics of irreversible processes. Addison-Wesley, Reading
52. Hahn W (1967) Stability of motion. In: Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen, Bd 138. Springer, Berlin
53. Haken H (1964) Z Phys 181:96
54. Haken H (1970) In: Encyclopedia of Physics, vol XXV/2c: Laser Theory. Springer, Berlin
55. Haken H (1975) Phys Lett 53A:77
56. Haken H (1975) Z Phys B 21:105, B 22:69, B 23:388
57. Haken H (1979) Light, vol 1, Elements of Quantum Optics. North-Holland Physics Publishing, Amsterdam, New York

58. Haken H (1985) Light, vol 2, Laser Light Dynamics. North-Holland Physics Publishing, Amsterdam, New York
59. Haken H (1996) Principles of brain functioning. Springer, Berlin
60. Haken H (2000) Information and Self-Organization, 2nd edn. Springer, Berlin
61. Haken H (2004) Synergetic computers and cognition, 2nd edn. Springer, Berlin
62. Haken H (2004) Synergetics: An introduction and advanced topics. Springer, Berlin
63. Haken H, Graham R (1971) Umschau 6:191
64. Haken H, Kelso S, Bunz H (1985) Biol Cybern 51:347
65. Haken H, Wunderlin A (1982) Z Phys B 47:179
66. Hansch D (2002) Evolution und lebenskunst. Grundlagen der psychosynenergetik. Ein Selbstmanagement-Lehrbuch. Vandenhoeck und Ruprecht, Göttingen
67. Hopf E (1942) Abzweigung einer periodischen lösung eines differentialsystems. Berichte der Mathematisch-Physikalischen Klasse der Sächsischen Akademie der Wissenschaften zu Leipzig XCIV 1, Leipzig
68. Hopf E (1948) Commun Pure Appl Math 1:303
69. Horsthemke W, Lefever R (1983) Noise-Induced transitions, Springer Series in Synergetics, vol 15. Springer, Berlin
70. Hoyle RG (2006) Pattern formation. Cambridge UP, Cambridge
71. Iooss G, Joseph DD (1980) Elementary stability and bifurcation theory. Springer, Berlin
72. Îto K (1969) Stochastic processes. Universitet Matematisk Institut, Aarhus
73. Jaynes ET (1957) Phys Rev 106:4, 620, Phys Rev 108:171
74. Jaynes ET (1967) In: Delaware Seminar in the Foundations of Physics. Springer, Berlin
75. Kadanoff LP, Götze W, Hamblen D, Hecht R Lewis EAS, Palcanskas VV, Rayl M, Swift J, Aspnes D, Kane J (1967) Rev Mod Phys 39:395
76. Kelley A (1967) In: Abraham R, Robbin J (eds) Transversal Mappings and Flows. Benjamin, New York
77. Kelso JAS (1981) Bull Psyconomic Soc 18:63
78. Kelso JAS (1995) Dynamic patterns: The self-organization of brain and behavior. MIT Press, Cambridge
81. Kernbach S (2008) Structural self-organization in multi-agent and multi-robotic systems, Thesis. Stuttgart University, Logos, Berlin
79. Kielhöfer, Hansjörg (2004) Bifurcation theory, an introduction with applications to PDES. Springer, Berlin
80. Kolmogorov AN (1954) Dokl Akad Nauk USSR 98:527
82. Kuhn TS (1996) The structure of scientific revolutions, 3rd edn. University of Chicago Press, Chicago
83. Kuramoto Y (1984) Chemical oscillations, waves and turbulence. Springer, Berlin
84. Kuznetsov, Yuri A (1995) Elements of applied bifurcation theory. Springer, Berlin
85. Köhler W (1920) Die physischen gestalten in ruhe und im stationären zustand. Vieweg, Braunschweig
86. Landau LD, Lifshitz IM (1959) In: Course of Theoretical Physics, vol 5, Statistical Physics. Pergamon Press, London-Paris
87. Landauer R (1975) Phys Rev A 12:636
88. Langevin P (1908) Sur la théorie du movement brownien. CR Acad Sci Paris 146:530
89. Lorenz EN (1963) J Atmospheric Sci 20:130, 20:448
90. Lyapunov AM (1906) Sur la masse liquide homogène donnée d'un movement de rotation. Zap Acad Nauk St. Petersburg 1:1
91. Ma, Tian, Wang, Shouhong (2005) Bifurcation theory and applications. World Scientific, Singapore
92. Manneville P (1990) Dissipative structures and weak turbulence. Academic Press, San Diego
93. Marx K (1987) Analytische und numerische behandlung der zweiten instabilität beim Taylor-Problem der flüssigkeitsdynamik, Thesis. Stuttgart University, Shaker, Aachen
94. May RM (1976) Nature 261:459
95. Meinhardt H (1982) Models of biological pattern formation. Academic, London
96. Meinhardt H (1990) The beauty of sea shells. Springer, Berlin
97. Meystre P, Sargent M (1990) Elements of quantum optics. Springer, Berlin
98. Mikhailov AS (1993) Foundations of synergetics. In: Distributed Active Systems, II (with AY Loskutov): Complex Patterns. Springer, Berlin
99. Mills DL (1991) Nonlinear optics basic concepts. Springer, Berlin
100. Moser J (1967) Math Ann 169:136
101. Murdock J (2002) Normal forms and unfoldings for local dynamical systems. Springer, New York
102. Murray JD (1989) Mathematical biology, 2nd edn 1993, 3rd edn 2002/2003. Springer, Berlin
103. Nayfeh, Ali H (1993) Method of normal forms. Wiley, New York
104. Nekorkin VI, Velarde MG (2002) Synergetic phenomena in active lattices. Patterns, waves, solitons, chaos. Springer, Berlin
105. Newell AC, Whitehead JA (1969) J Fluid Mech 38:279
106. Newhouse S, Ruelle D, Takens F (1978) Commun Math Phys 64:35
107. Nicolis G (1995) Introduction to nonlinear science. Cambridge University Press, Cambridge
108. Nicolis G, Prigogine I (1977) Self-organization in non-equilibrium systems. Wiley, New York
109. Pauli H (1928) Probleme der modernen physik. In: P Debye (ed) Festschrift zum 60. Geburtstag A Sommerfelds. Hirzel, Leipzig
110. Pismen LM (1999) Vortices in nonlinear fields. Clarendon Press, Oxford
111. Pismen LM (2006) Patterns and interfaces in dissipative dynamics. Springer, Berlin
112. Pliss VA (1964) Izv Akad Nauk SSSR, Mat Ser 28:1297
113. Poincaré H (1960) Les methods nouvelles de la méchanique céleste. Gauthier-Villars, Paris. Reprint 1892/99 Dover Publ, New York
114. Prigogine I, Nicolis G (1967) J Chem Phys 46:3542
115. Rabinovich MI, Ezersky AB, Weidmann PD (2000) The dynamics of patterns. World Scientific, Singapore
116. Risken H (1965) Z Phys 186:85
117. Risken H (1989) The Fokker Planck eq. Springer, Berlin
118. Ruelle D, Takens F (1971) Commun Math Phys 20:167
119. Sargent M, Scully MO, Lamb WE (1974) Laser physics. Addison-Wesley, Reading
120. Sattinger DH (1980) Group theoretic methods in bifurcation theory. Lecture Notes Math, vol 762. Springer, Berlin
121. Schawlow AL, Townes CH (1958) Phys Rev 112:1940
122. Schiepek G (1999) Die grundlagen der systemischen therapie. Theorie-Praxis-Forschung. Vandenhoeck und Ruprecht, Göttingen

123. Schleich WP (2001) Quantum optics in phase space. Wiley-VCH, Weinheim
124. Schmidt E (1908) Zur theorie der linearen und nichtlinearen integralgleichungen, Teil 3. Math Annalen 65:370
125. Schöll E (2001) Nonlinear spatio-temporal dynamics and chaos in semiconductors. Cambridge University Press, Cambridge
126. Scully M, Lamb WE (1967) Phys Rev 159:208 (1968) 166:246
127. Segel LA (1969) J Fluid Mech 38:203
128. Sell GR (1979) Arch Ration Mech Anal 69:199
129. Shannon CE, Weaver W (1949) The mathematical theory of communication. Univ of Illin Press, Urbana
130. Smale S (1967) Bull AMS 73:747
131. Sounders PT (1980) An introduction to catastrophe theory. Cambridge University Press, Cambridge
132. Sparrow CT (1982) The Lorenz equations: bifurcations, chaos and strange attractors. Springer, Berlin
133. Staliunas K, Sanchez-Morcillo V, Gaul LJ (2003) Transverse patterns in nonlinear optical resonators. Springer, Berlin
134. Stratonovich RL (1963) Topics in the theory of random noise, vol 1, 1967 vol II. Gordon Breach, New York-London
135. Swift J, Hohenberg PC (1977) Phys Rev A 15:319
136. Swinney HL, Gollub JP (eds) (1981) Hydrodynamic instabilities and the transition to turbulence, Topics Appl Phys, vol 45. Springer, Berlin
137. Taylor GI (1923) Phil Trans R Soc Lond A 223:289
138. Thom R (1975) Structural stability and morphogenesis. Benjamin, Reading
139. Turing AM (1952) Phil Trans R Soc Lond B 237:37
140. Vavilin VA, Zhabotinsky AM, Yaguzhinsky LS (1967) Oscillatory processes in biological and chemical systems. Moscow Science Publ, Moscow, p 181
141. Walgraef D (1997) Spatio-temporal pattern formation. Springer, New York
142. Walls DF, Milburn GJ (1994) Quantum optics. Springer, Berlin
143. Weidlich W (2000) Sociodynamics. A systematic approach to mathematical modelling in the social sciences. Harwood Academic Publishers, Amsterdam
144. Weidlich W, Haake F (1965) Z Physik 186:203
145. Wigner EP (1932) Phys Rev 40:749
146. Wilson KG, Kogut J (1974) Phys Rep 12 C:75
147. Wolpert L (1969) J Theor Biol 25:1
148. Wunderlin A, Haken (1981) Z Phys B 44:135
149. Zaikin AN, Zhabotinsky AM (1970) Nature 225:535

Books and Reviews

Springer Series in Synergetics. Founded by Haken H (1977) vols 1–ca 100. Springer, Berlin

Synergetics, Introduction to

HERMANN HAKEN
Institut für Theoretische Physik, Universität Stuttgart,
Stuttgart, Germany

Synergetics (Greek **synergeon**: science of cooperation) is an interdisciplinary field of research that deals with the behavior of complex systems, i. e. systems composed of many individual parts that in general may produce complex behavior. The systems considered may belong to a great variety of fields, ranging from natural sciences with physics, chemistry, biology, through medicine and psychology to economy, ecology, management theory etc. The central topic of the synergetics enterprise is this: By means of their cooperation, the individual parts of a system can spontaneously produce structures, i. e. special arrangements between the elements or specially coordinated behavior without specific steering from the outside, i. e. without the interference of an external agent. The physical systems, studied by synergetics, are away from thermal equilibrium. The structures may be spatial patterns, be it regular or irregular, temporal patterns, i. e. all kinds of self-sustained oscillations, ranging from harmonic oscillations to chaotic behavior, spatio-temporal patterns, or functional patterns such as produced, for instance, by the brain of humans or animals, giving rise to specific behavior. The basic question synergetics asks is this: Are there general principles that govern the self-organization of systems, irrespective of the nature of the individual parts? When this question was posed some 40 years ago by H. Haken, the problem put forward seemed to be absurd in view of the great variety of possible elements. However, this research program has turned out to be quite successful as witnessed, for instance by the Springer Series in Synergetics, as well as by an independent book series in Russian. The price to be paid for approaching this goal is as follows:

1. Focus the study on those situations where the **macroscopic state** of a system changes **qualitatively**.
2. Start from comparatively simple systems that are provided either by systems in physics (e. g. lasers or fluids) or model systems dealing with self-organization processes in a variety of fields, such as biology.

As has become evident over the past few decades, self-organization phenomena occur in a vast variety of fields whose presentation actually would fill a whole encyclopedia, which would have led to an encyclopedia within an encyclopedia. Indeed the reader him- or herself may find numerous examples in his/her own field of research. In view of this situation, I felt it would be wise to present to the reader a few basic concepts and only a few prototypical examples. It is, indeed, an outstanding fact that few concepts allow us to cover a great variety of self-organization phenomena from a unifying point of view. Among these concepts are stability, instability, control parameters, order parameters and the slaving principle. Future research will have to add more concepts as has already happened, for instance, in chaos theory.

As the reader will note, from the theoretical point of view, synergetics has become a meeting place between bifurcation theory and mathematical theories related to it, the theory of stochastic processes, and phase transition theory, actually in the sense of the Landau theory. The mathematical tools employed are generalized Langevin equations and the Fokker–Planck equation, and to some extent also the density matrix equation. In physics, it has become possible to start from first principles. For instance in quantum optics, based on Heisenberg equations of motion for operators, the coherence properties of laser light (and other non-linear optical devices) were derived in every detail (its explicit presentation would require, however, a whole handbook article, so that the reader is referred to the original literature). Fluid dynamics was chosen as a highly illustrative and prototypical example of pattern formation in systems away from thermal equilibrium. When a fluid is energetically excited, with an increasing degree of excitation it may run through a hierarchy of spatio-temporal patterns. At comparatively low excitation levels, well defined patterns evolve. While regular patterns stand in the foreground of the article ► Fluid Dynamics, Pattern Formation by Bestehorn, the article by Pismen ► Patterns and Interfaces in Dissipative Dynamics emphasizes defects and interfaces. At higher excitation levels, we reach the field of turbulence, which even after the ground-breaking work by Kolmogorov is still a hot subject of research. Important steps are done here, e. g. by the study of the dynamics of vortices. The work by Friedrich and Peinke, authors of the article ► Fluid Dynamics, Turbulence comprises also the Fokker–Planck equation approach. Actually, in the introductory article ► Synergetics: Basic Concepts by Haken, a method using the concept of information (entropy) is outlined how to derive a Fokker–Planck equation from measured data. Friedrich (private communication) and Peinke found a more direct access, the results of which are briefly outlined in the article by Friedrich and Peinke.

A quite modern line of research on complex systems follows up the method of the Fokker–Planck equation and especially of **non-linear** Fokker–Planck equations. This will be outlined in the article ► Linear and Non-linear Fokker–Planck Equations by Frank.

So far, all these articles deal with "hard" science, especially physics. As it has turned out over past decades, the general concepts of synergetics have great potentialities in fields often called "soft" science such as movement science, medicine, and even psychology and psychiatry. Actually, as the following articles will show, basic concepts of synergetics help to convert soft science into hard science. The article by Fuchs and Kelso ► Movement Coordination deals with movements of humans (and animals). But though these "systems" are highly complex, large classes of transitions between movement coordination can be theoretically and experimentally treated in great detail, even of a prototypical character.

While these approaches might be called "macroscopic", the article by Tass, Popovych and Hauptmann ► Brain Pacemaker penetrates into the microscopic level of brain functions, namely by modeling the collective behavior of neurons and their reactions to external interventions. This paper is remarkable, because it shows a new aspect of the research on coupled nonlinear oscillators. Whereas so far the problem of how such systems synchronize, was in the foreground of research, now for medical reasons, namely to fight Parkinsons's disease, the problem arises how to desynchronize such a system by appropriate interventions.

The mathematical theory of synergetics or perhaps in a wider sense dynamical systems theory with concepts such as attractors, allows one to cast earlier concepts developed by Gestalt theory into clear-cut mathematical models. The order parameters just play the role of Gestalt. Synergetics may be considered as a theory of indirect control of systems, in that it provides psychiatrists with insights for how to intervene with their clients. A comprehensive article, dealing with these aspects, is given in ► Self-Organization in Clinical Psychology by Schiepek and Perlitz. This article shows how methods of mathematical analysis are penetrating more and more into the field of psychiatry, and allowing doctors to monitor the mental and behavioral state of a patient.

In conclusion, in particular with respect to biological systems (but not exclusively), the following remarks are in order:

1. While in physical systems, at least those considered in the article on synergetics, the control parameters are fixed from the outside, in biological systems the control parameters are in general produced by the system itself. They are the slowly varying variables in this case. When their dynamics is admitted, specific phenomena may occur, such as the recognition of ambivalent figures (switching between two or several percepts).
2. In biological and some physical processes, pattern formation (morphogenesis) is a two- or multi-step process. In the first step **dynamic structures** are formed. In a second step these dynamic patterns are transformed into solid patterns. A nice example is provided by radiolarians in which eventually dynamic concentration

patterns of chemicals are transferred into skeletons by means of calcification. The exploration of such two- or multi-step processes is, as it seems, just at its beginning.

In the article ► Intentionality: A Naturalization Proposal on the Basis of Complex Dynamical Systems, W. Tschacher shows how mental processes can be linked to material processes where he tackles the longstanding problem of intentionality and relates it to a gradient dynamics which can be formalized by concepts of synergetics.

Last but not least, Synergetics as a theory dealing with systems composed of many individual parts deals with the interrelation between human individuals and their ability to form specific social groups, viewed as a process of self-organization. An outstanding example is provided by Juval Portugali's article ► Self-Organization and the City.

System Dynamics, Analytical Methods for Structural Dominance Analysis in

Christian Erik Kampmann[1], Rogelio Oliva[2]

[1] Department of Innovation and Organizational Economics, Copenhagen Business School, Copenhagen, Denmark

[2] Mays Business School, Texas A&M University, College Station, USA

Article Outline

Glossary

Behavior mode The traditional meaning of the term is the qualitative nature of the observed system behavior, such as damped or expanding oscillations, overshoot and collapse, exponential growth or adjustment to equilibrium, or limit cycles. In linear systems theory, the term has a more specific meaning, cf. the explanation for eigenvalues.

Bode plot (phase and gain plot) A tool used in classical control theory to characterize the frequency response, i. e., the amplification A and phase shift ϕ in the system output variable of interest $x(t) = A \sin(\omega t + \phi)$ compared to the input variable $u(t) = \sin(\omega t)$, as a function of the frequency ω of the input.

Chaos A type of behavior exhibited by nonlinear systems that appears to be approximately periodic but with a seemingly random element. A hallmark of chaotic behavior is that it is sensitive to initial conditions.

Dominant structure A general term for the feedback loops (or possibly external driving forces) that are "most important" in generating a behavior pattern of interest. In nonlinear models, particularly single-transient models, there is frequently a shift in structural dominance, i. e. in the strength and significance of certain feedback loops.

Dynamic decomposition weights (DDW) An application of Eigenvector Elasticity Analysis (EVA) that focuses on how parameter changes influence the relative weights (DDW's) of the system behavior modes in a particular variable.

Eigenvalue An eigenvalue for a square matrix $\boldsymbol{A}$ is a value λ for which the equation $\boldsymbol{Ar} = \lambda \boldsymbol{r}$ has a non-zero solution $\boldsymbol{r} \neq \boldsymbol{0}$. The column vector $\boldsymbol{r}$ is called the (right) eigenvector corresponding to the eigenvalue λ. The eigenvalues and eigenvectors determine the behavior modes (components) in the solution to the linear dynamical system $\dot{\boldsymbol{x}} = \boldsymbol{Ax}$. A real eigenvalue λ leads to an exponential behavior mode $\exp(\lambda t)$ while a complex eigenvalue $\lambda = \tau \pm i\omega$ leads to oscillatory behavior modes $\exp(\tau\, t) \sin(\omega t + \phi)$. The eigenvectors determine the weight, or the degree to which a particular behavior mode is expressed in a particular system variable.

Eigenvalue elasticity analysis (EEA) A method of analyzing the significance of a structural element, say a loop or a link in the model with a gain g, in terms of its marginal effect upon the eigenvalues λ of the system. There are several such measures, such as the *influence measure* $\partial\lambda/\partial g \cdot g$, the *elasticity* $\partial\lambda/\partial g \cdot (g/\lambda)$, or, in the case of complex-valued eigenvalues, the effect upon the damping ratio, natural frequency, damping time, etc., as illustrated in Fig. 7. See also *Loop Eigenvalue Elasticity Analysis (LEEA)*.

Eigenvector See explanation for *Eigenvalue*.

Eigenvector elasticity analysis (EVA) A complement to Eigenvalue Elasticity Analysis (EEA) that looks explicitly at the expression or relative weight of each behavior mode in each system variable. These weights are related to the eigenvectors of the system matrix.

Frequency domain A term used to describe the analysis of signals with respect to frequency. While a time domain graph shows the behavior of the signal over time, the frequency domain graphs shows how much of the signalvariance lies within each given frequency band.

Independent loop set (ILS) Although the number of feedback loops in a model can be very large (theoretically astronomically large), there is a much smaller *independent loop set* that can be considered independent structural elements. For a strongly connected system (where any pair of variables are connected via causal chain in both directions) with N links and n variables, there are exactly $N - n + 1$ independent loops. Simple algorithms exist for constructing independent loop sets, in particular *Shortest Independent Loop Sets (SILS)*. See also explanation for *Loop Eigenvalue Elasticity Analysis (LEEA)*.

Linear dynamical system A system where the rates $\dot{\boldsymbol{x}} = (\mathrm{d}x_1/\mathrm{d}t, \ldots, \mathrm{d}x_n/\mathrm{d}t)$ are a linear function of the state variables $\boldsymbol{x} = (x_1, \ldots, x_n)$ and exogenous or control variables $\boldsymbol{u} = \left(u_1, \ldots, u_p\right)$, expressed by the equation $\dot{\boldsymbol{x}} = \boldsymbol{Ax} + \boldsymbol{Bu}$ where $\boldsymbol{A}$ is an $n \times n$ matrix and $\boldsymbol{B}$ is an $n \times p$ matrix. Unlike nonlinear systems of the general form $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u})$, linear systems have analytical solutions based on the eigenvalues and eigenvectors of the matrix $\boldsymbol{A}$ (cf. explanation for *Eigenvalues*).

Linear systems theory The mathematical theory of *linear dynamical systems*.

Loop eigenvalue elasticity analysis (LEEA) A form of eigenvalue elasticity analysis (EEA) that uses graph theory to express structural changes in terms of change in the strength of individual feedback loops. *Independent* loops can be assigned individual (loop) eigenvalue elasticities or influence measures just like other structural elements (see explanation for *Eigenvalue Elasticity Analysis (EEA)* and *Independent Loop Set (ILS)*).

Model simplification approach A way of attributing dynamic behavior to particular pieces of structure by replacing the full model with a simplified structure. See also *Structure contribution approach*.

Nonlinear systems Systems of the form $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u})$ where $\boldsymbol{f}$ is a nonlinear function. See explanation for *Linear dynamical systems*.

Pathway participation metric A measure that decomposes the curvature ($\ddot{x} = \mathrm{d}^2x/\mathrm{d}t^2$) of a variable x_i into the individual driving components, $\ddot{x}_i = \sum_j \partial\dot{x}_i/\partial x_j \cdot \dot{x}_j$. By considering the sign of the curvature relative to the slope, i. e., $\ddot{x}/\dot{x}$, one may define behavior as (apparently) dominated by positive $\ddot{x}/\dot{x} > 0$ or negative $\ddot{x}/\dot{x} < 0$ feedback loops. The component (pathway) with the largest absolute value and the same sign as $\ddot{x}/\dot{x}$ is then defined as the dominant structure.

Quasilinear models Models that are almost linear in structure around the operating point of interest so that they may be well approximated by a linear model.

Quasiperiodic behavior A behavior that is a sum of oscillations of incommensurate frequencies so that the system never returns to exactly the same point (which would be the case for periodic behavior).

Shortest independent loop set (SILS) An *Independent Loop Set (ILS)* that consists of the shortest possible loops (in terms of the number of nodes and links in each loop). Since the choice of ILS is far from unique, an SILS provides a more focused choice of loops, which are typically also easier to interpret due to their short length.

Single-transient models Models where the behavior of interest is the transition toward an equilibrium or constant growth rate. Models are typically nonlinear, exhibit patterns such as smooth transition, overshoot and collapse, growth, or stagnation.

Structure contribution approach A way of linking model structure to dynamic behavior by considering how individual pieces of structure (feedback loops or subsystems) contribute to the behavior pattern of interest by turning the structure on or off (in traditional simulation experiments) or by considering the marginal effect of small changes in structure (the eigenvalue approach). See also *Model simplification approach*.

Definition of the Subject

The link between system structure and dynamic behavior is one of the defining elements in the system dynamics paradigm, yet it is only recently that systematic, mathematically rigorous methods for exploring this link have started to become available. In a sense, a simulation model can be viewed as an explicit and consistent theory of the behavior it exhibits. Although this point of view has certain merits, not least the fact that it lifts the discussion from outcomes to causes of these outcomes and from events to underlying structure [11,59], we are concerned here with a more compact explanation of the system's behavior. In fact, most system dynamics modeling projects report their results in terms of simpler explanations of the observed results, typically in terms of dominant feedback loops that produce the salient features of the behavior.

Most often, dominant structure is thought of in terms of feedback loops and, occasionally, external driving forces to the system. For simple systems with relatively few variables it is usually easy to use intuition and trial and error simulation experiments to explain the dynamic behavior as resulting from particular feedback loops. In larger systems, this method becomes increasingly difficult and the risk of incorrect explanations rises accordingly. There is a need, therefore, for analytical methods that provide some consistency and rigor to this process.

These analytical tools are important to the practitioner because the structure-behavior link is the key to finding leverage points for policy initiatives. And they are important to the theorist because a system dynamics theory of a particular phenomenon is an account of how certain feedback loops cause certain dynamic patterns of behavior to appear. The qualitative understanding of the model behavior is often at least as important as the particular numerical predictions obtained, even in applied studies. Yet the rigor of such an account depends directly on the rigor with which structure-behavior link can be made in a given model.

The classical disciplines of linear systems theory and control engineering have provided a set of concepts and tools, particularly system eigenvalues and eigenvectors, that can also be applied under many circumstances to the nonlinear models found in system dynamics, not as a complete theory but as a pragmatic aid. This article reviews the recent advances in analytical tools based on linear systems theory and discusses its future potential for the both the system dynamics practitioner and the theorist.

Though we strongly believe in the utility of these methods, it is important to realize that advances in nonlinear dynamics and complexity theory in recent decades have shown that it is not possible to construct a complete theory of dominant structure because nonlinear systems are capable of exceedingly complex and intricate behavior that is impossible to predict without actually simulating the system. Furthermore, applications of graph theory to system dynamics models have revealed that the concept of feedback loops has some inherent problems and limitations because there are potentially many different loop descriptions of the same system (see [28,40]). Thus, the analytical tools should be viewed as pragmatic aids to model analysis that can guide the modeler's intuition, rather than universal methods that provide automatic answers.

We first provide a brief historical introduction to the different ways scholars have thought about the notion of dominant structure, including an example of the traditional approach to structural analysis. In the next section we present the formal mathematical representation of linear and nonlinear systems and how one may describe the dynamic behavior in terms of behavior modes and system eigenvalues. In the four following sections we present alternative approaches to performing this analysis. We conclude with a summary of the current state of research and a discussion of future directions.

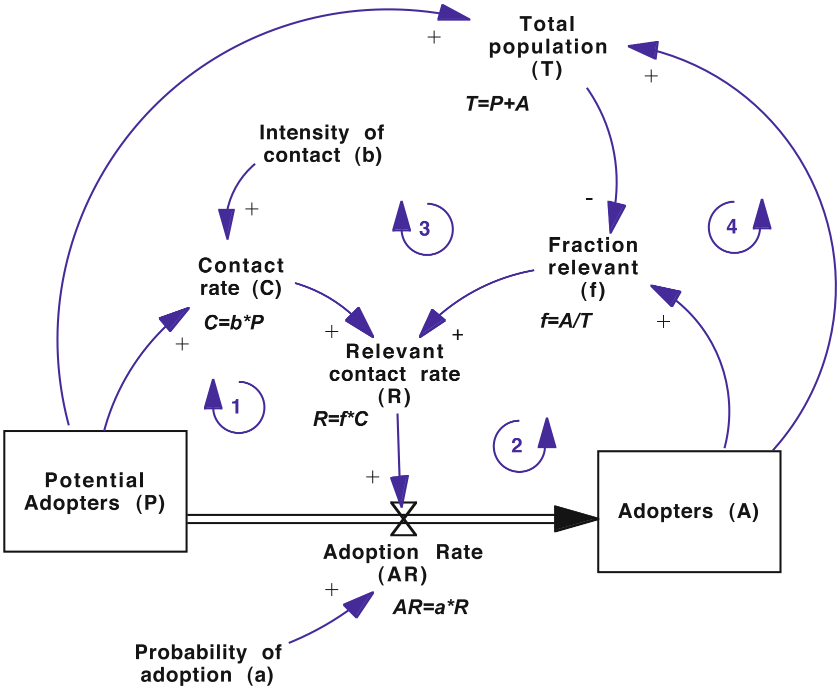
Introduction

Understanding model behavior is closely related to the process of model testing and validation, for which there is a well-established tradition and an extensive literature in the field (e. g., [2,10,17,36,46,47]). Indeed there is no sharp line between model building, testing, validation, and analysis – in practice, the analyst undertakes all these processes simultaneously [17].

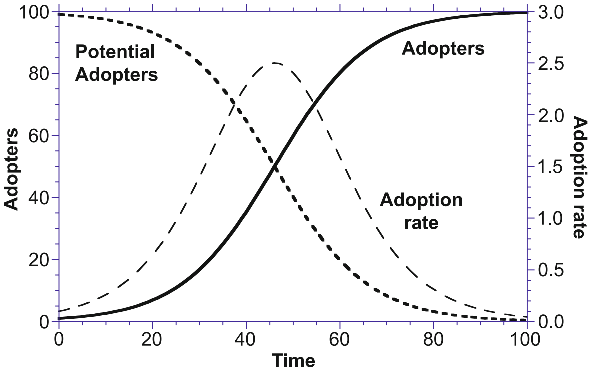
Of particular concern is whether one can identify pieces of structure that are in some sense "important" in generating the observed behavior of interest. Traditionally, system dynamics analysts have relied on trial-and-error simulation to discover these structures, by changing parameter values or switching individual links and feedback loops on and off. The tradition is well developed and includes a set of principles for partial model formulation and testing based the organizational theory of bounded rationality [27,36].

The intuition guiding this effort often relies on simple feedback systems with one or a few state variables, where the behavior is fully documented and understood. In particular, the modeler uses well-understood "generic structures" that seem to appear again and again in system dynamics models, such as "overshoot and carrying capacity collapse", "drifting goal structure", etc. (see [30,56,60] for an account of these structures). Clearly such structures can be a useful aid to understanding if the model is sufficiently simple to allow such simple structures to be identified.

A simple example of a generic structure is the classical model of diffusion, sometimes known as the Bass model ([3], see also Chapter 9 in [59]). The model structure is illustrated in Fig. 1, and the resulting behavior, an s-shaped growth curve, is illustrated in Fig. 2. The idea behind the model is that the adoption of a new technology is driven by the number of users that have already adopted it, through a word-of-mouth effect. One may interpret the s-shaped behavior as the interaction of two feedback loops, namely loop no. 2, the positive "word-of-mouth", and loop no. 1, the negative "exhaustion" loop (see Fig. 1). In the beginning, the positive loop dominates, leading to exponential growth in the number of adopters. Later, however, the negative loop gains strength, and the behavior shifts to an exponential adjustment toward the eventual market saturation. Thus, the traditional feedback loop analysis helps give an intuitive understanding of the dynamics of the model.



System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 1
The Bass model of diffusion



System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 2
Behavior of the Bass model

In large-scale models with perhaps hundreds of state variables, however, the traditional approach shows significant limitations. In practice, model building and analysis is often done using a "nested" partial model testing approach where one goes from the level of small pieces of structure to entire subsystems of the model, with frequent re-use of known formulations and partial models. Although this approach does carry a long way, it can be very difficult to discover feedback mechanisms that transcend model substructures in ways not anticipated by the modeler in the original dynamic hypothesis. Thus, there is a danger that observed behavior is falsely attributed to certain feedback mechanisms when in fact another set of feedbacks is driving the outcome. Likewise, one may make false inferences about how a particular feedback mechanism modifies the behavior, e. g., whether it attenuates or amplifies a particular oscillation.

Modern software packages can run extensive tests for sensitivity and "reality checks" where a large number of parameters are varied simultaneously [44]. This is clearly a significant improvement over "manual" trial and error methods, particularly when these methods are combined with statistical inference methods such as Kalman Filtering or Monte Carlo maximum likelihood estimation [6,8,39,43,45,55]. A variant of this approach involves using statistical experimental design and correlation methods to screen for significant model structure (parameters), as suggested by Ford and Flynn [9]. Indeed, the prospects of marrying such methods with modern search and optimization methods like classifier systems [26] or genetic algorithms [19] seem very promising. However, these methods are more addressing issues in estimation, validation and testing than inferences about or understanding how (dominant) structure is causing behavior.

Richardson [47] suggested a taxonomy of approaches to the notion of dominant structure, where he distinguishes along three dimensions, namely linear vs. nonlinear systems, model reduction vs. loop contribution, and the characterization of behavior in terms of time graphs vs. eigenvalues or frequency response. Of these, the distinction between model reduction and loop contribution is the most important.

In the model reduction approach, the idea is to replace a large complicated model with a simplified smaller model that captures the "essence" of the dynamics. A good example of this is Sterman's simple model of the economic long wave [58], which was distilled from the much larger System Dynamics National Model [18]. Eberlein [5,7] attempted to tackle model simplification in a systematic way in linear systems by focusing on retaining specific behavior modes. In large part his results were negative: it is generally not possible to build simpler models that reproduce the salient behavior without sacrificing either the accuracy of the behavior or the ability to relate the simplified model variables to those in the full model. It is fair to say that this line of inquiry has largely been abandoned as a result. Extracting the "essence" of a model remains an art more than a science.

The focus here will be on Richardson's second category, the loop contribution or, more generally, the *structure contribution* approach. It reflects the intuitive idea that if one removes the element under consideration, e. g. by weakening a link or switching off a feedback loop, and the behavior then "disappears", one would say that the element in some sense "causes" the observed behavior.

This notion underlies the traditional trial-and-error simulation approach, sometimes supplemented with methods from the classical control engineering, which focuses on how structural elements modify the behavior of the system, viewed in terms of the frequency response. Typically, the method works "backwards" by starting with simple feedback systems of single loops and then considering the marginal effect of adding links and loops. We discuss this approach in Sect. "Traditional Control Theory Approaches" below.

If, instead, one considers marginal (infinitesimal) changes in structure, e. g. in the strength of a particular link, it is possible to derive rigorous analytical results for the resulting change in behavior expressed as the eigenvalues of the linearized model. One would then say that if a change in a system element has a relatively large effect upon the behavior pattern of interest, this element is "significant" in "causing" the behavior. This is what underlies the *eigenvalue elasticity* and *eigenvector* approaches discussed in Sects. "Eigenvalue Elasticity Analysis", "Eigenvectors and Dynamic Decomposition Weights (DDW)". The marginal and experimental approaches may supplement each other well, where a marginal analysis may identify elements that can then be tested experimentally for their significance.

Unlike the traditional control method and the eigenvalue method that work in the structural and *frequency domain*, the *pathway participation* method (PPM) relates directly to the time path of particular system variables and is more concerned with the qualitative nature of the time path, expressed in terms of signs of the slope (whether growing or declining) and curvature (whether convex or concave) than with numerical measures of degree of influence. Briefly stated, the PPM traces the causal links in the variables influencing the system variable in question and then identifies the most important chain of links. We discuss this method in Sect. "Pathway Participation Metrics".

Common to the approaches discussed here is that they all build upon a precise mathematical characterization of the system behavior. In the next section, we demonstrate how the concepts from linear systems theory may be used to give a precise characterization of behavior in terms of component *behavior modes*.

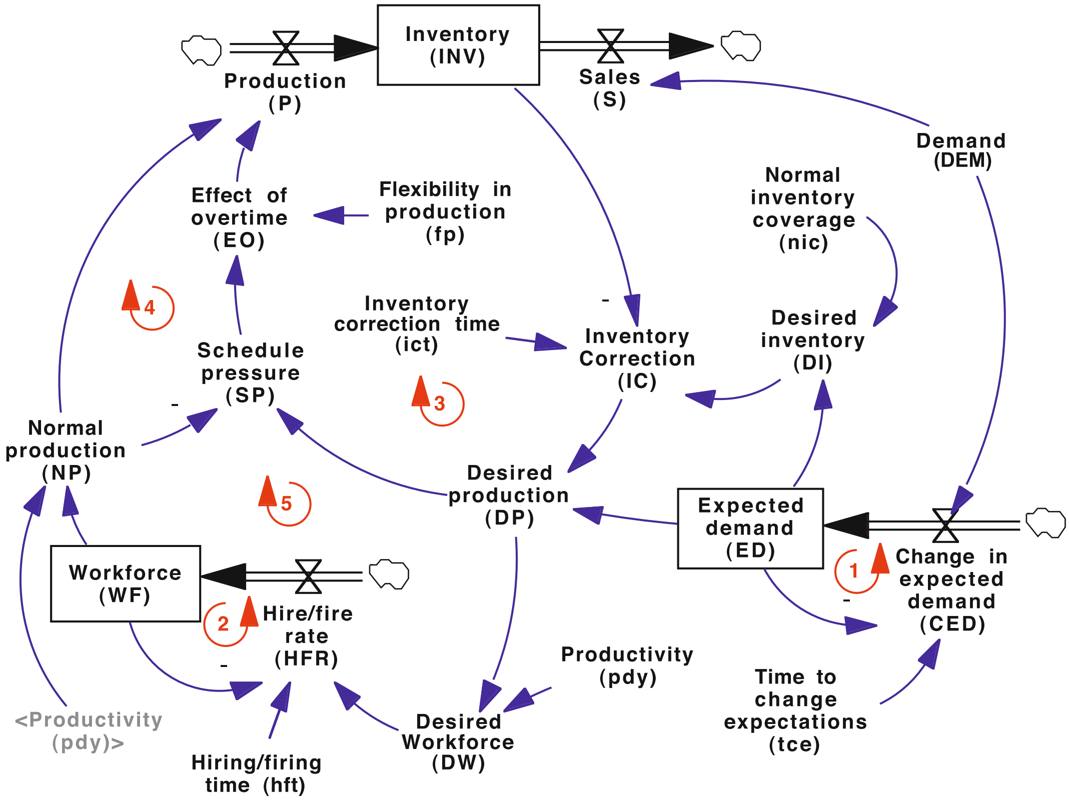
Characterizing Linear and Nonlinear Systems

A system dynamics model can be represented mathematically as a set of ordinary differential equations

$$\frac{\mathrm{d}\boldsymbol{x}(t)}{\mathrm{d}t} \equiv \dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{u}(t)) , \qquad (1)$$

where $\boldsymbol{x}(t)$ is a (column) vector of n state variables (levels) $(x_1(t), \ldots, x_n(t))$, $\boldsymbol{u}(t)$ is a column vector of p exogenous variables or control variables $\left(u_1(t), \ldots, u_p(t)\right)$, $\boldsymbol{f}()$ is a corresponding vector function, and t is simulated time. In this paper, we restrict our attention to the state variables (levels) of the model for notational convenience, ignoring the auxiliary variables. Mathematically, a model can always be brought to the *reduced* form (1), but in practice, the auxiliary variables give a more intuitive account of the analysis. Likewise, we do not consider time-varying systems (where time t enters as an explicit argument in the function $\boldsymbol{f}$), since these can usually be accommodated by an appropriate definition of the exogenous variables $\boldsymbol{u}$. In general, $\boldsymbol{f}$ is a nonlinear function of its arguments, and we speak of a *nonlinear* system. Conversely, if $\boldsymbol{f}$ is a linear function, we speak of a *linear* system.

Figure 3 and Table 1 show a well-known example, the inventory–workforce model. It has three state variables, Inventory (INV), Workforce (WF), and Expected Demand



System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 3

Flow diagram of the inventory workforce model

(ED), and one exogenous variable, Demand (DEM), i. e.,

$$\boldsymbol{x}(t) = \begin{pmatrix} \text{INV} \\ \text{WF} \\ \text{ED} \end{pmatrix} ; \quad \boldsymbol{u}(t) = (\text{DEM}) , \qquad (2)$$

and the function $\boldsymbol{f}$ is determined by the equations in Table 1.

Given the model structure (1), knowledge of the initial conditions $\boldsymbol{x}(0)$, and the path of the input variables $\boldsymbol{u}(t)$, the behavior of the model is completely determined. It is in this sense that the model structure (1) constitutes a "theory" of the time behavior $\boldsymbol{x}(t)$, as mentioned in the introduction. Yet, we are interested in methods that yield a more compact explanation, short of having to simulate the entire model structure.

It turns out that in its ultimate form, this dream is beyond reach: Since the days of Henri Poincaré, mathematicians have known that it is impossible to find general analytical solutions to nonlinear systems. Furthermore, the development of nonlinear dynamics and chaos theory has proven that such systems, even when they have very few state variables, can produce highly complex and intricate behavior that goes beyond general analytic methods (e. g., [42,48]). Thus, we will never find a final general theory where we can infer the behavior of the system directly from its structure; instead, we will always have to rely on simulation to discover the dynamics implied by the structure. (This is not to say that no general analytical results exist in nonlinear systems. The field of chaos theory has uncovered a number of universal features, e. g., relating to the transition from periodic or quasi-periodic behavior to chaos, where the transitions show both qualitative and quantitative similarities that are independent of the specific forms of the model equations (see, e. g., [42]). However, these universal features relate to specific situations such as period-doubling or intermittency routes to chaos).

The best we can hope for, therefore, is a set of tools that will guide intuition and help identify *dominant structure* in the model. By dominant structure we mean particular feedback loops that are in some sense "important" in shaping the behavior of interest. To the extent that we can identify such dominant structures, we may say that we have found a "theory" of the observed behavior.

Although the term "behavior" may appear rather loose, experience and reflection tells us that there is a limited number, perhaps a dozen or so, of relevant behav-

System Dynamics, Analytical Methods for Structural Dominance Analysis in, Table 1
Equations of the inventory workforce model

Equation	Name	Units
d/dt(ED) = (DEM − ED)/tce, ED_0 = DEM * df	Expected demand	[Units/Month]
d/dt(INV) = P − S, INV_0 = DI	Inventory	[Units]
d/dt(WF) = HFR, WF_0 = DWF	Workforce	[Workers]
S = DEM	Shipments	[Units/Month]
P = NP * EO	Production	[Units/Month]
EO = fp * (1 − (1 − 1/fp)^SP)	Effect of overtime	[Dimensionless]
NP = WF * pdy	Normal production	[Units/Month]
DI = ED * nic	Desired inventory	[Units]
SP = DP/NP	Schedule pressure	[Dimensionless]
DEM = 1(Exogenous)	Demand	[Units/Month]
DP = ED + IC	Desired production	[Units/Month]
IC = (DI − INV)/tci	Inventory correction	[Units/Month]
HFR = (DWF − WF)/hft	Hire/fire rate	[Workers/Month]
DWF = ED/pdy	Desired workforce	[Workers]
hft = 5	Hire/fire time	[Month]
pdy = 1	Productivity	[Units/Month/Worker]
tce = 4	Time to change expectations	[Month]
fp = 1.05	Flexibility in production	[Dimensionless]
nic = 3	Normal inventory coverage	[Months]
ict = 2	Inventory correction time	[Months]
df = 0.5	Disequilibrium fraction	[Dimensionless]

ior patterns that dynamical systems can exhibit. Some of these behaviors, like exponential growth, exponential adjustment, and damped or expanding oscillations, are typical of linear systems. Others, like limit cycles, quasiperiodic motion, mode-locking, and chaos, can only be exhibited by nonlinear systems.

Common to the approaches considered in this paper is that they are based on tools from linear systems theory, i. e., they approximate the nonlinear model (1) with a linearized version, using first-order Taylor expansion around some operating point $\boldsymbol{x}_0, \boldsymbol{u}_0$, i. e.,

$$\dot{\boldsymbol{x}}(t) \approx \boldsymbol{f}(\boldsymbol{x}_0, \boldsymbol{u}_0) + \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}(\boldsymbol{x} - \boldsymbol{x}_0) + \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}}(\boldsymbol{u} - \boldsymbol{u}_0) , \quad (3)$$

or, by redefinition of the variables $\boldsymbol{x} \to \boldsymbol{x} - \boldsymbol{x}_0 - \boldsymbol{f}(\boldsymbol{x}_0, \boldsymbol{u}_0) \times (t - t_0)$ and $\boldsymbol{u} \to \boldsymbol{u} - \boldsymbol{u}_0$,

$$\dot{\boldsymbol{x}}(t) \approx \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{u}(t) , \quad (4)$$

where $\boldsymbol{A}$ is constant $n \times n$ matrix of partial derivatives $\partial f_i / \partial x_j$ and $\boldsymbol{B}$ is constant $n \times p$ matrix of partial derivatives $\partial f_i / \partial u_j$, and all partial derivatives are evaluated at the operating point.

For the linear system (4), there is a well-developed and extensive theory of the system behavior as a function of its structure, expressed in the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. One may broadly distinguish two parts of the theory, named classical control theory (e. g., [38]) and modern linear systems theory (e. g., [4,31]). We return to the classical control theory in the next section.

Modern control theory or linear systems theory (LST) is concerned with the dynamical properties of the system as a direct function of the system matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. A key element in this theory is the notion of the system *eigenvalues*, i. e., the eigenvalues of the matrix $\boldsymbol{A}$. If, for simplicity, we restrict ourselves to the endogenous dynamics of the system (set $\boldsymbol{u} = 0$), we can write the solution to (4) as

$$x_i(t) = c_{i,1} \exp(\lambda_1 t) + c_{i,2} \exp(\lambda_2 t) + \cdots + c_{i,n} \exp(\lambda_n t) , \quad i = 1, \ldots, n , \quad (5)$$

where $\lambda_1, \ldots, \lambda_n$ are the n eigenvalues of the matrix $\boldsymbol{A}$ and $c_{i.j}$ are constants that depend upon the eigenvectors and the initial condition of the system. In other words, the resulting behavior is a weighted sum of distinct *behavior modes*, $\exp(\lambda t)$. If an eigenvalue is real, the corresponding behavior mode is exponential growth (if $\lambda > 0$) or exponential decay (if $\lambda < 0$). Complex-valued eigenvalues come in complex conjugate pairs $\lambda = \tau \pm i\omega$ which give rise to oscillations $\exp(\tau\, t) \sin(\omega t + \phi)$ of frequency ω that are either expanding (if $\tau > 0$) or damped (if $\tau < 0$).

In this manner, the eigenvalues serve as a compact and rigorous characterization of the behavior (of linear systems).

At any point in time, any system, linear or nonlinear, may be approximated by the expression (5). Whether it remains a good approximation depends upon how much and how quickly the eigenvalues change due to the nonlinearities in the function $\boldsymbol{f}$. If they are more or less constant for significant periods of time, we may speak of *quasilinear systems* that are well approximated by the linear system. In some cases, however, the eigenvalues change so rapidly that it makes little sense to characterize the behavior by equation (5). (See [29] for further discussion).

Traditional Control Theory Approaches

The first set of methods, which we call the traditional approach, has been used for decades and is part of the standard curriculum in system dynamics teaching at the graduate level. It involves using the concepts from classical control theory [38] to very simple systems with only a few state variables.

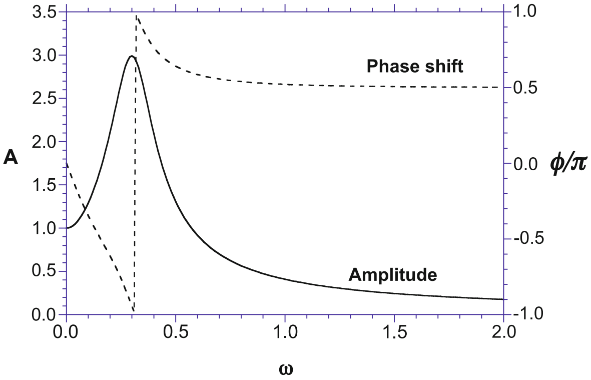
The starting point is the simple first- and second-order positive and negative feedback loops found in any introductory treatment of system dynamics. The advantage of the approach is its simplicity. Although it serves at a guide to intuition, however, the obvious shortage is that it applies rigorously only to simple systems. There have been some attempts to treat higher-order systems by adding a few feedback loops [23], but the step to large-scale models is beyond this method given its inherent limitations.

Graham [23] distills a number of principles that are based on the metaphor of a "disturbance" traveling along the chain of causal links in a feedback loop and getting amplified, damped, and possibly delayed in the process. For major negative feedback loops, which are known to tend to produce oscillation, adding minor negative loops and cross-links, or shortening the delay times increases the damping. Conversely, adding positive loops in to the oscillatory system tends to lengthen the period of oscillation whereas the effect on the damping depends upon the delays in the positive loop. Using the metaphor of pushing a child on a swing, it becomes clear that the timing of the propagation of a disturbance has as much importance for its effect on the damping as its strength.

For analyzing the behavior of positive feedback loops, Graham suggested calculating the Open-loop steady-state gain (OLSSG), a measure of the amplification around the loop. A gain greater than unity will result in exponential growth while gains less than 1 will give exponential adjustment (leveling off or decay). The intuition is perhaps best illustrated by an example: sales-driven growth. Suppose a salesperson can eventually pull in \$100,000 per month in orders (probably with a several-month long delay), and assume that the company allocates 10% of revenue to marketing. Then this eventually leads to \$10,000 per month for sales efforts. If the cost of a salesperson (salary, overhead, expenses etc.) is, say, \$8,000 dollars per month, then the efforts of the current sales force will provide enough revenue to support $10,000/8,000 = 1.25$ persons per current person. Thus, the OLSSG of the positive loop from salespersons → orders → revenues → marketing budget → salespersons is 1.25, and the system will grow exponentially (until other factors limit the growth). Conversely, if the gain is less than 1, one salesperson will not sell enough to support their own cost, and the loop will lead to exponential decay. Graham showed how the actual rate of growth is partly determined by the OLSSG, and partly by the time constants (delays etc.) involved. (See also Subsect. 15.3 in [59]).

In the context of oscillating systems, system dynamics has also employed a concept from classical control theory, frequency response. The frequency response is determined from the transfer function of the system, $G(i\omega)$, which is a complex-valued function that specifies how an input signal $u(t)$ with frequency ω results in an output signal $x(t)$ that may be phase shifted (delayed), and either amplified or attenuated. For linear systems, G can be calculated directly from the system matrices in (4) – the transfer function (matrix) is $\boldsymbol{G}(i\omega) = \boldsymbol{B}(i\omega\boldsymbol{I} - \boldsymbol{A})^{-1}$, where $\boldsymbol{I}$ is the identity matrix (see e. g. [4]). For nonlinear systems, G may be found through simulation experiments.

Usually, G is represented in a *Bode* or *phase-and-gain* diagram. For instance, Fig. 4 shows a Bode diagram of the



System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 4
Phase-and-gain diagram (Bode diagram) showing the inventory $A\sin(\omega t + \phi)$ with amplitude A and phase shift ϕ, relative to a sinusoidal demand $\sin(\omega t)$, for varying values of the frequency ω of the demand fluctuation

inventory variable INV (t) relative to the exogenous demand input variable DEM (t) in the inventory workforce model in Fig. 3. The diagram shows how the relative amplitude of the oscillation and the relative phase shift (in radians) between input and output varies as a function of the frequency of the input.

It is clear from the diagram that there is a certain frequency range, around the system's own natural frequency, where fluctuations in demand are greatly amplified compared to other frequencies. Indeed, it is a general phenomenon in systems that they will tend to amplify certain frequencies while attenuating other frequencies. This may be used to explain or understand the role of particular structures in the model in generating oscillation at certain frequencies, even when there are no oscillations coming in from the outside world. (External random noise is enough to produce oscillations in the system because random noise contains fluctuations at all frequencies). In this manner, the approach nicely demonstrates the "endogenous viewpoint" that behavior (oscillations) is generated internally by the system. As an analytic tool for large scale systems, however, the method does not seem to produce any additional insights. Thus, we may conclude that the classical approaches serve mostly as intuitive metaphors to guide the analyst rather than as full analytical tools.

Pathway Participation Metrics

The *pathway participation* method [34,35] represents a further development of an original suggestion by Richardson [46] to provide a rigorous definition of loop polarity and loop dominance. Richardson motivated this with the common confusion associated with positive feedback loops, which may exhibit a wide range of behaviors [23], as Barry Richmond noted with wonderful humor:

"Positive loops are … er, well, they give rise to exponential growth … or collapse … but only under certain conditions … Under other conditions they behave like negative feedback loops …" [49].

Richardson proposed that the polarity of a loop be defined as the sign of the expression

$$\frac{\partial \dot{x}_i}{\partial x_i} = \frac{\partial f_i\,(\boldsymbol{x}\,,\boldsymbol{u})}{\partial x_i}\,, \tag{6}$$

in the model (1), with a positive sign indicating a positive loop and vice-versa. When several loops operate simultaneously, the sign of the expression indicates whether the positive or negative loops dominate. Note, however, that the definition only applies to minor loops (i. e. loops involving a single level). Put differently, it only considers the diagonal elements of the matrix $\boldsymbol{A}$ in the linearized system (4). Richardson [46] demonstrates how even with this limitation, analyzing the system with this metric can (sometimes) yield insights into behavior of higher-order systems.

The expression (6) hints that it is relevant to consider the curvature, i. e., the second time derivative, $\ddot{x}$, of a variable when looking for dominant structure. Although he does not say so explicitly, this is effectively the focus of Mojtahedzadeh's pathway method. Figure 5 shows how one may classify behavior by comparing the first and second time derivatives of a variable. As seen in the figure, the sign of the expression $\ddot{x}/\dot{x}$, which Mojtahedzadeh denotes the total *pathway participation metric* or *PPM*, indicates whether the behavior appears dominated by positive or negative loops, much in line with Richardson's definition of dominant polarity. A zero curvature indicates a shift in loop dominance (cf. the middle column in the figure). Note, however, that the interpretation of the middle row in the figure where the slope $\dot{x}$ is zero has no clear interpretation in terms of loop dominance. Indeed this hints at one of the weaknesses of the approach that we will return to below.

Mojtahedzadeh's method proceeds by decomposing the PPM into its constituent terms as follows,

$$\mathrm{PPM}_i = \frac{\ddot{x}_i}{\dot{x}_i} = \sum_{j=1}^{n} \frac{\partial f_i}{\partial x_j}\frac{\dot{x}_j}{\dot{x}_i}\,, \tag{7}$$

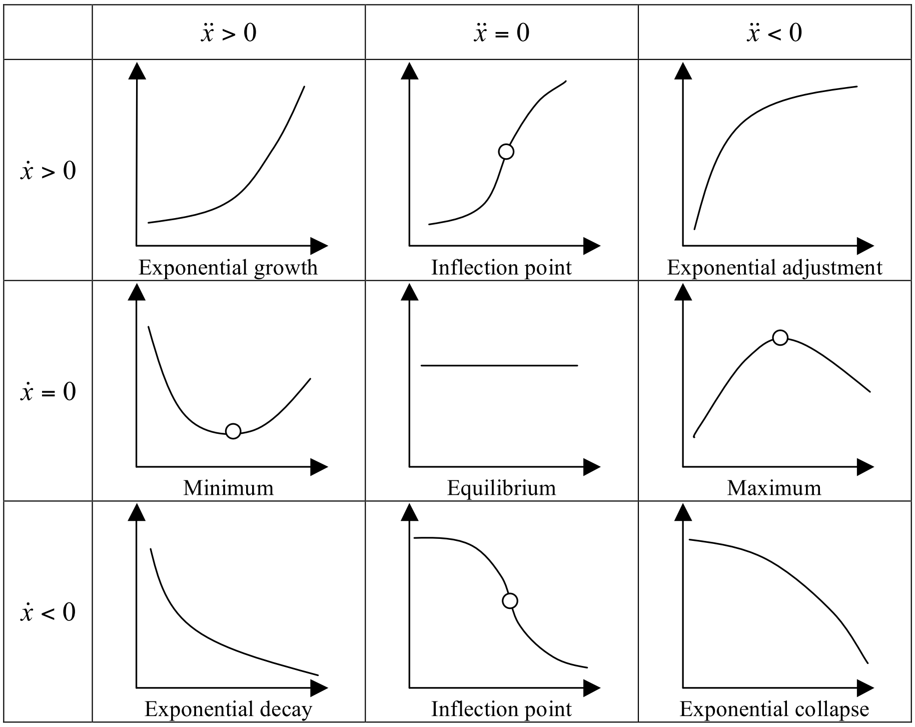
where, for brevity, we have chosen to ignore the exogenous variables u. One might say that each of the terms in the sum in (7) represents the separate influence of each of the systems' state variables on the behavior of x_i. Mojtahedzadeh in fact uses a normalized measure for the terms,

$$\frac{\left(\partial f_i/\partial x_j\right)\dot{x}_j}{\sum_{k=1}^{n}\left|\left(\partial f_i/\partial x_k\right)\dot{x}_k\right|}\,, \tag{8}$$

which can vary between -1 and $+1$, to measure the relative importance of the pathway from variable j. By explicitly considering auxiliary variables y in the model, one may further decompose each term $\partial f_i/\partial x_j$ into a sum of terms

$$\frac{\partial f_i^k}{\partial x_j} = \frac{\partial f_i}{\partial y_1}\cdot\frac{\partial y_1}{\partial y_2}\cdot\cdots\cdot\frac{\partial y_{m-1}}{\partial y_m}\cdot\frac{\partial y_m}{\partial x_j}\,, \tag{9}$$

corresponding to a causal chain or pathway $\pi_k = \{x_j \to y_m \to \cdots y_2 \to y_1 \to \dot{x}_i\}$. Mojtahedzadeh now considers each possible pathway (9) and defines the "dominant" pathway as the one with the largest numerical value



System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 5
Characteristic behavior patterns based on the first and second time derivatives

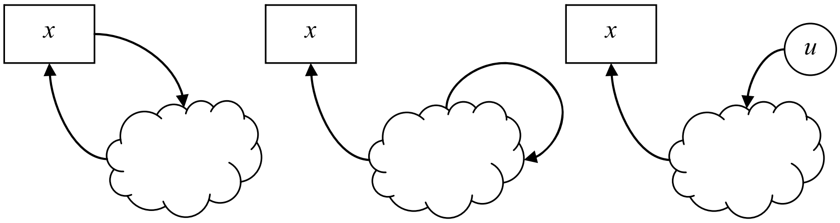
and the same sign as PPM_i. Having selected this dominant pathway, $\pi^*_{ij} = \{x_j \to y_m \to \cdots y_2 \to y_1 \to \dot{x}_i\}$, which originates in the state variable x_j the procedure is repeated for that state variable x_j, and so forth, until one either reaches one of the already "visited" state variables (in which case a loop has been found) or an exogenous variable (in which case an external driving force has been found). Thus, the procedure may result in three alternative forms of dominant structure illustrated in Fig. 6, namely a "pure" minor or major feedback loop, a pathway from a feedback loop elsewhere in the system, or a pathway from an exogenous variable.

By dividing the observed model behavior into different phases according to the taxonomy in Fig. 5 and then applying the method just described at different points in during these phases, one can reveal how the dominant structure changes over time. For illustration, the PPM method is applied to the Bass model and the results are presented in Fig. 7. The figure shows the metrics of four alternative pathways (four feedback loops) and the results accord nicely with the informal analysis done earlier: The method identifies two phases, exponential growth, exponential adjustment, and identifies the "word-of-mouth" positive loop (loop 1) as dominant in the first phase and the "exhaustion" loop (loop 2) as dominant in the second phase.

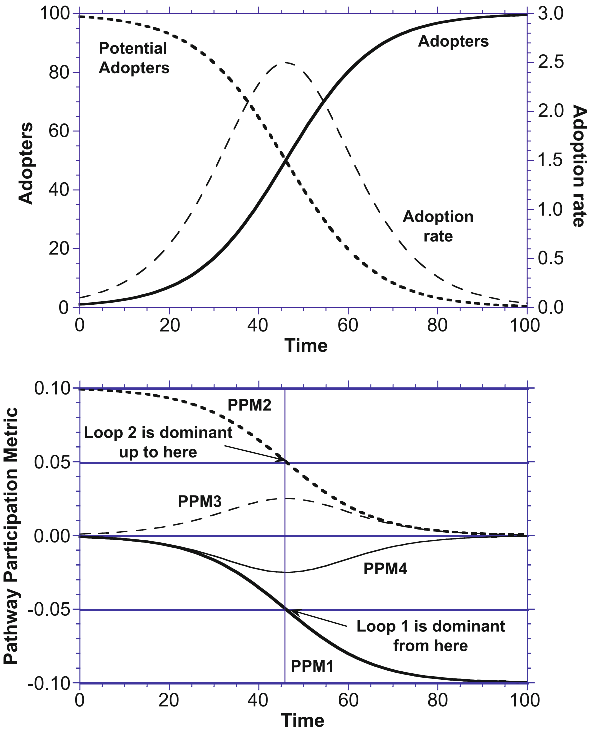
The PPM method is still mostly used at an early explorative stage on rather simple models, where it does appear to aid insight into the dynamics (e. g. [41]), and has been implemented in a software package, *Digest*, [35].

From the studies performed so far, it is clear that the main strength of this method is its relative computational simplicity (it does not require computing eigenvalues, which is a numerically demanding task), and the intuitive and direct connection it makes between the observed behavior and the influencing structural elements. Unlike the other approaches which operate in the "frequency domain", the method considers the time path of a specific variable directly.

There are, however, some important outstanding issues that remain to be clarified. First, the method is not suitable for oscillatory systems. The problem is easy to recognize when one considers how the PPM measure will vary over the course of a sinusoidal outcome: The sign of the PPM will shift twice during each cycle, indicating that the behavior is alternately dominated by positive and negative loops, even though the system structure, and hence the loop dominance, may remain unchanged all the time.



System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 6
Three alternative forms of dominant structure in the PPM method



System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 7
Pathway participation measures in the Bass model

Richardson [46] already alluded to this problem by noting that the measure only considers the diagonal elements in the system matrix in (4), yet we know that the structure causing oscillation is the major negative loop that involve the off-diagonal elements. This is a significant limitation, given the prevalence and importance of oscillation in system dynamics analysis.

A second limitation of the current implementation of PPM is that it uses a depth-first search for the single most influential pathway for a variable. This strategy does not capture the situation where more than one structure may contribute significantly to the model behavior and, through the depth-first algorithm, may miss alternative paths that could prove to yield a larger total value of the metric. This problem could be addressed by modifying the search algorithm and is most likely of minor importance.

Another issue is how to treat the case when $\dot{x} = 0$ since it appears in the denominator of the terms in (6). However, it is not clear that it is necessary to do this division, given that it is easy to identify the nine cases in the figure by simply examining its sign. Thus, the issue is probably not of much significance.

The fourth issue, on the other hand, is more significant, namely the emphasis on identifying a single “dominant” structure. In reality, of course, the behavior of a variable is influenced by many loops and pathways at once. Reducing the consideration to a single one of these may miss important features of the structure-behavior relationships. For instance, a variable may be influenced by two negative loops and one positive, with the sum of the two negative loops dominating the influence of the positive loop, even though that loop by itself has the strongest influence on the behavior. It is more appropriate to consider the relative importance of alternative pathways, yet the method does not address how one would partition the behavior among pathways (the three structures in Fig. 6) – only among individual links.

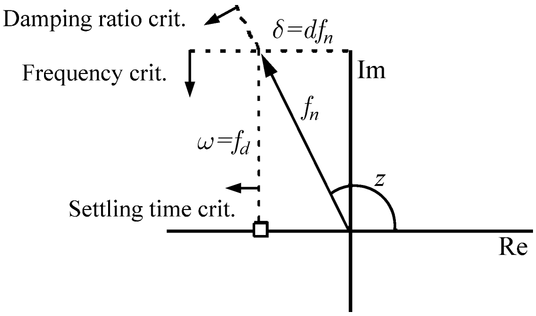
Thus, while the notion of pathways seems an interesting and useful idea, it may be that it will ultimately be more effective to use a list, ranked in order of magnitude, of the pathways that influence a variable.

Finally, the method shares a weakness with the traditional method in that it considers primarily partial system structures rather than global system properties. In contrast, the two eigenvalue methods to which we now turn are based on a rigorous characterization of the entire system (at a given point in time).

Eigenvalue Elasticity Analysis

The third method may be termed *eigenvalue elasticity analysis* (or EEA for short) and builds upon the tools from modern linear systems theory (LST), applied to the linearized model (4). The method is concerned with the structural elements that significantly affect the system eigenvalues or behavior modes – the values λ in (5). Specifically, it measures influence by the elasticity of an eigenvalue λ with respect to some parameter g in the model, defined as $\varepsilon = \left(\partial\lambda/\partial g\right)(g/\lambda)$, i. e. the fractional change in the eigenvalue relative to the fractional change in the parameter. The advantage of this fractional measure is that it is dimensionless, i. e., independent upon the choice of units, including the time scale unit. Sometimes, the influence measure is used instead, defined as $\mu = (\partial\lambda/\partial g)g$. This measure has dimension $[1/time]$ and so depends upon the choice of is time unit, but it is generally easier to interpret for complex-valued eigenvalues and avoids numerical problems with very small or zero eigenvalues (see [29,54]).

The idea behind EEA was first introduced in system dynamics by Forrester [14] in the context of economic stabilization policy. For purposes of policy analysis in oscillating systems, one may define a number of criteria from engineering control theory, all of which relate to the eigenvalues of the system, as summarized in Table 2. Figure 8 provides a graphical characterization of the eigenvalues and policy criteria in the complex plane. Though these measures are not new, the EEA method is unique in its attempt to use them to gain qualitative intuitive understanding of the system. A significant step in this direction was first suggested by Forrester [15] with the notion that the elasticities of any links in the model (corresponding to elements of the matrix $\boldsymbol{A}$ in the linearized system (4)), can be interpreted as the sum of elasticities of all feedback loops containing that link. We have chosen to name this approach *loop eigenvalue elasticity analysis (LEEA)*.



System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 8
Characterization of eigenvalues plotted in the complex plane

Kampmann [28] provided a rigorous definition of LEEA and also pointed to the fact that feedback loops are not independent. In other words, given the possibly very large number of loops in a given model (Kampmann demonstrated how the theoretical maximum number of loops grows combinatorically with the number of variables), it only makes sense to speak of individual contributions of a limited set of *independent* loops. He proved that a fully connected system (where there is a feedback loop between any pair of variables – the typical case in system dynamics models) with N links and n variables has a total of $N - n + 1$ independent loops and provided a procedure for constructing this set and calculating the loop elasticities.

Kampmann's analysis points to a fundamental issue relating to the notion of feedback loops as a way to explain

System Dynamics, Analytical Methods for Structural Dominance Analysis in, Table 2
Stabilization policy criteria and corresponding effects on eigenvalues and BDW of a policy change in a system element *g*

Policy Criterion	Description	Change in eigenvalue $\lambda = \delta \pm i\omega, \omega > 0$	Change in BDW w	Appropriate measure in time path
Damping	Increases the rate of decay of oscillation (or decreases the rate of expansion)	$\frac{\partial\delta}{\partial g}\frac{g}{\delta} < 0$	N/A	$\frac{x(t+T)}{x(t)}$
Frequency	Decreases the frequency of oscillation (or lengthens the period *T*)	$\frac{\partial\omega}{\partial g}\frac{g}{\omega} < 0$	N/A	T
Variance	Reduces the variance of a target variable (or the weighted average variances of several variables)	No simple relation	$\frac{\partial w}{\partial g}\frac{g}{w} < 0$	$\int x(t)^2\,\mathrm{d}t$
Auto-spectrum	Reduces variance of target variable(s) within a target frequency range	No simple relation	$\frac{\partial w}{\partial g}\frac{g}{w} < 0$	Filter in frequency domain
Frequency response gain	Reduces the gain (amplification) in the target frequency range for a particular combination of disturbance exogenous and output variables.	Based upon transfer function $G(i\omega)$		

System Dynamics, Analytical Methods for Structural Dominance Analysis in, Table 3

Loops and their influences in the inventory workforce model. Values are measured at time $t = 0$. The model contains three eigenvalues, $\lambda_1 = -0.250$ and $\lambda_2, \lambda_3 = -0.138 \pm i\,0.285$. The influence measure is defined as $g \cdot \partial\lambda/\partial g$. For the imaginary part, a positive influence measure means that the frequency is increased

Loop	Nodes	Gain	Influence on Re[λ_1]	Influence on Re[λ_2]	Influence on Im[λ_2]
1	ED > CED	−0.250	−0.250	0.000	0.000
2	W > HFR	−0.200	0.000	−0.100	−0.022
3	INV > IC > DP > SP > EO > P	−0.076	0.000	−0.038	0.008
4	INV > IC > DP > DW > HFR > W > NP > P	−0.100	0.000	0.000	0.176
5	INV > IC > DP > DW > HFR > W > NP > SP > EO > P	0.015	0.000	0.000	−0.027

behavior: the significance assigned to a particular loop depends upon the context (the chosen independent loop set). In other words, feedback loops are derived and relative concepts rather than fundamental independent building blocks of systems. Oliva [40] further refined the definition of independent loop sets by introducing the *Shortest independent loop set (SILS)* along with a procedure for constructing the set. Although a SILS is not generally unique, experience seems to suggest that it is easier to interpret [41]. Yet the issue remains that independent feedback loop sets are relative concepts.

In Table 3, we show how the LEEA analysis applies to the simple inventory–workforce model in Fig. 3. The model contains a total of 5 feedback loops, all of which are independent. The loops are listed in Table 3, including their constituent variables (nodes), and the gain of the loop (defined in a similar manner to the pathway participation metrics above). We see that there are three minor negative loops, related to the exponential smoothing of expected demand (loop 1) and the adjustment of workforce to desired workforce (loop 2). The minor loop 3 is the "overtime shortcut" that allows production to adjust part way to desired production immediately so one does not have to wait for the workforce to adjust. Loop 4 is the main major negative loop that adjusts inventory to desired levels via workforce adjustment. Finally, loop 5 (the only positive loop) is a fairly weak loop that moderates the effect of loop 4 by adjusting the overtime effect "back to normal" when the workforce is brought in line with desired production.

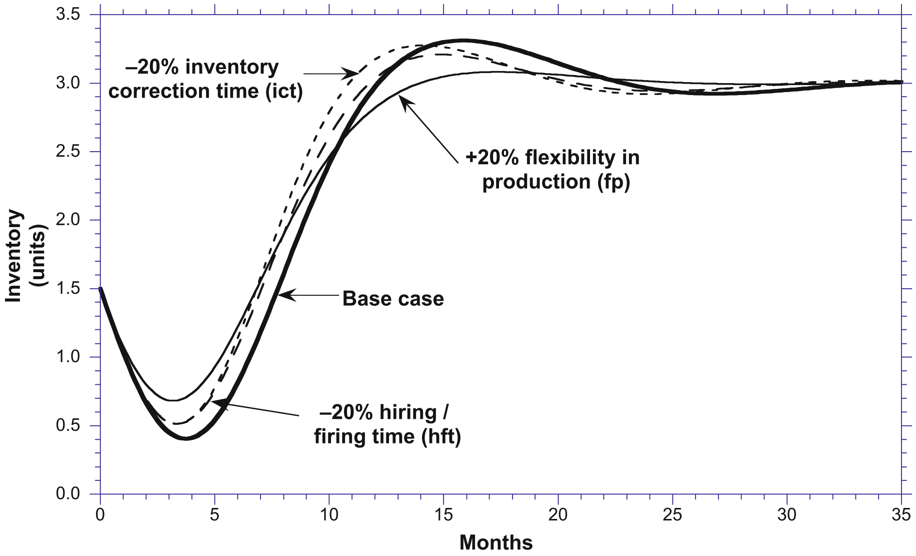
Although the model is nonlinear (due to the overtime function), the eigenvalues do not change very much over the course of its behavior. The model contains one real eigenvalue ($\lambda_1 = -0.250$) and one pair of complex conjugate eigenvalues ($\lambda_2, \lambda_3 = -0.138 \pm i\,0.285$). The first eigenvalue corresponds to the adjustment of expected demand (ED). The other pair produces a damped oscillation in inventory and workforce.

Table 3 also shows the loop influences upon the three eigenvalues. Note how there is a one-to-one correspondence between loop 1 (the adjustment of expected demand) and the first eigenvalue. This is due to fact that the ED level constitutes a single strongly connected component of the model (see Fig. 3), i. e. there is no feedback between this level and the rest of the model. We also note that the workforce adjustment and the overtime loops have a stabilizing influence upon the behavior (they make the real part of the oscillatory eigenvalues more negative and have relative little effect upon the frequency of oscillation). Conversely, the major negative loop 4 has a destabilizing influence, since strengthening it will increase the frequency of oscillation and not increase the damping. The effects of loop 5 are fairly weak.

From this analysis, one would therefore expect parameters that strengthen loop 2 (shortening hire/fire time) or loop 3 (increase overtime effect) would stabilize the system while strengthening loop 4 (shorter inventory adjustment time) will destabilize the system. Indeed this is what happens, as illustrated in the simulations in Fig. 9.

The EEA/LEEA method has been applied in a number of contexts (e. g. [1,20,22,24,29,51,52,54]), but remains a tool employed only by specialists in fundamental research, not least because it has not been incorporated into standard software packages. Thus, the potential of the method for widespread practice remains unexplored.

One might be skeptical that a method derived from linear systems theory may have any use for the nonlinear models found in system dynamics. Kampmann and Oliva [29] considered what types of models the method would be particularly suited for. They defined three categories of models, based upon the behavior they are designed to exhibit: 1) linear and quasilinear models, 2) nonlinear single-transient models, and 3) nonlinear periodic models. The first category encompasses models of oscillations, possibly combined with growth trends, with relatively stable equilibrium points, (e. g., the classical in-



System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 9
Simulated behavior of inventory–workforce model, showing the effect on inventory of parameter changes for overtime (flexibility of production), inventory adjustment time (ict), and labor hiring/firing time (hft), respectively

dustrial dynamics models [11]). Nonlinearities may modify behavior (particularly responses to extreme shocks) but the instabilities and growth trends can be analyzed in terms of linear relationships. Kampmann and Oliva concluded that LEEA showed the most promise and potential for this class of models because the analytical foundations are solid and valid, and because the method has the ability to find high-elasticity loops even in large models very quickly without much intervention on the part of the analyst.

The second class is typical of scenario models like the World Model [13,33], the Urban Dynamics Model [12], or the energy transition model in [57], to name a few, that show a single transient behavior pattern, like overshoot and collapse or a turbulent transition to a new equilibrium. In these models, nonlinearities usually play an essential role in the dynamics. Yet it is possible to divide the behavior into distinct phases where certain loops tend to dominate the behavior. In this class of models LEEA also shows promise by measuring shifts in structural dominance by the change in elasticities. But it requires more input from the analyst (e. g. in defining the different phases of the transition) and it has no obvious advantage over other methods, like PPM.

The third class, nonlinear periodic models, are those that exhibit fluctuating behavior in which nonlinearities play an essential role, such as like limit cycles, quasiperiodic behavior, or chaos, (see, e. g. [48]). Here the utility of the method is much less clear and depends upon the specifics of the model in question. For example, the classic Lorenz model that exhibits limit cycles, period doubling and deterministic chaos does not lend itself to any insight using LEEA [29]. This is particularly the case in systems with strong nonlinearities such as min and max functions. In these systems, the behavior may change abruptly (eigenvalues suddenly shift) in what is called border-collision bifurcations [37,61]. In other cases, the method of breaking the behavior into phases with different dominant structures may yield significant insight from LEEA. For instance, Sterman's simple long wave model [58] lends itself well to this approach (e. g. [24,28]).

In the present paper, we add a fourth category of models or behavior for which the method has not been explored yet. We name this category nonlinear multi-modal models. These encompass the cases where one behavior mode interacts with and therefore modifies another behavior mode – something that can only happen in nonlinear systems. The most common example is mode-locking or entrainment, in which oscillations become synchronized (e. g. [25]). Another example is mode modification, where one behavior mode (growth or oscillation) affects the character of another (typically oscillation). An example of this is the interaction of the business cycle with the economic long wave, where the former tends to get more severe during long wave downturns [16]. Whether LEEA can contribute to this class of models remains to be seen.

Compared to the former two methods, the EEA/LEEA is mathematically more general and rigorous, though many of the mathematical issues in the method remain to be addressed, as we summarize below. This rigor is also the main strength of the method, since it provides an unambiguous and complete measure of the influence of the entire feedback structure on all behavior modes.

A weakness or challenge that is starting to show up is the computational intensity in calculating eigenvalues and elasticities. This is not so much an issue of computer time and memory space as of the stability of numerical methods. Kampmann and Oliva [29] found that the numerical method used sometimes proved unstable, yielding meaningless results. Clearly, there is a need to explore this issue further, possibly building upon the developments in control engineering.

A more serious weakness is the difficulty in interpreting the results: Eigenvalues do not directly relate to the observed behavior of a particular variable. The concepts of eigenvalues and elasticities are rather abstract and unintuitive [10]. There is a need for tools and methods that can translate them into visible, visceral, and salient measures. Here, the measures in Table 2 may provide a guide. In particular, it is possible to use (linear) filtering in the frequency domain to define a behavior of interest. For example, an analyst may be concerned with structures causing a typical business cycle (3–4-year oscillation) and, by specifying a filter that "picks out" that range of fluctuation, could obtain measures for structures that have elasticities in that range. Because filters are typically linear operators, all the analytical machinery of the LEEA method will also apply in this case – a significant advantage.

Using filters will also solve an issue that appears in large-scale models, namely the presence of several identical or nearly identical behavior modes. Saleh et al. [54] do consider the analytical problems associated with repeated eigenvalues, where it becomes necessary to use generalized eigenvectors, and where other behavior modes appear involving power functions of time. A filter essentially constitutes a weighted average of behavior modes and in this fashion avoids the "identity problem" of non-distinct eigenvalues.

The most serious theoretical issue, in our view, is how the results are interpreted using the feedback loop concept. As mentioned, the concept is relative (to a choice of an independent loop set). Moreover, practice reveals that the number of loops to consider is rather large and that the loops elasticities often do not have an easy or intuitive explanation. A lot of care must be taken when interpreting the results. For instance, Kampmann and Oliva [29] found that "phantom loops" – loops that cancel each other by logical necessity and are essentially artifacts of the equation formulations used in the model – could nonetheless have large elasticities and thus seriously distort the interpretation of the results. An example of "phantom loops" is found in the Bass model in Fig. 1, where loops 3 and 4 are artifacts of the way the model is formulated. If the variable *Total population* (T) was eliminated from the equations, the loops would disappear and in fact they exactly cancel each other out (since T is constant). Nonetheless, they appear on the list of loops and appear to have a separate influence on behavior. These kinds of problems may not be intractable, but their resolution will require careful mathematical analysis.

Finally, a problem with EEA and LEEA is that it only considers changes to behavior modes, not the degree to which these modes are expressed in a system variable of interest. This issue is addressed by also considering the eigen*vectors* of the system, which is the foundation for the analysis in the next section.

Eigenvectors and Dynamic Decomposition Weights (DDW)

The last set of methods, which are still in early development, we have termed the *eigenvector-based* approach (EVA). EVA attempts to improve the EEA/LEEA method by considering how much an eigenvalue or behavior mode is expressed in a particular system variable. The logic of the method and how EEA and EVA complement each other is shown in Fig. 10. As shown by Kampmann [28], in a sense there is a one-to-one correspondence between eigenvalues and loop gains whereas the eigenvectors arise from the remaining "degrees of freedom" in the system. The observed behavior of the state variables in the model is then the combined outcome of the behavior modes (from the loop gains) and the weights for each mode (from the eigenvectors) in the respective state variable.

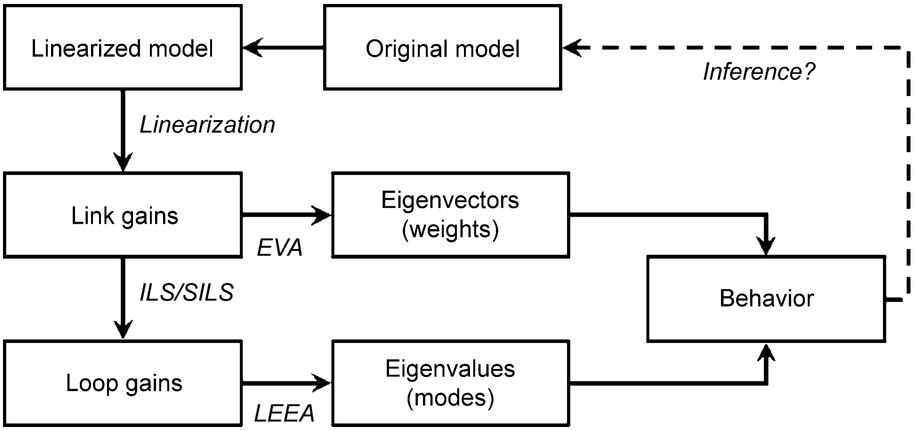
A number of researchers have attempted to develop EVA methods. Some emphasize the curvature (second time derivative) of the behavior, similar to the starting point of the PPM method [24,50,51,52]. The slope or rate of change $\dot{x}(t)$ of a given variable x in the linearized system may be written by

$$\dot{x}(t-t_0) = w_1 \exp(\lambda_1 (t-t_0)) + \cdots + w_n \exp(\lambda_n (t-t_0)) \,, \quad (10)$$

where the weights w_i are related to the eigenvectors. Then the curvature at time t_0 is

$$\ddot{x}(t_0) = w_1\lambda_1 + \cdots + w_n\lambda_n \,. \quad (11)$$

One may therefore interpret (11) has the sum of contribution from individual behavior modes. Güneralp [24]



System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 10
Schematic view of eigenvalue and eigenvector analysis approach

suggested using the terms on the right-hand side of (11) as weights to combine elasticities of individual behavior modes ε_i with respect to some system element (like a link gain or a loop gain) into a weighted sum

$$\bar{\varepsilon} = \frac{\sum_{i=1}^{n} w_i \lambda_i \varepsilon_i}{\sum_{i=1}^{n} |w_i \lambda_i|}\,, \tag{12}$$

as a measure of the overall significance of that system element. He further normalized the elasticity measure by the elasticity measure for other system elements, i. e., assuming there are K such elements (loops or links), the relative importance ρ_k of the kth element is defined as

$$\rho_k = \frac{\bar{\varepsilon}_k}{\sum_{j=1}^{K} |\bar{\varepsilon}_j|}\,, \tag{13}$$

with the motivation that elasticities may vary greatly in numerical values, making comparisons at different points in time difficult, whereas ρ_k is a relative measure varying between +1 and −1. His results shed an alternative light on the behavior of these models, but the mathematical meaning, consistency and significance of the doubly normalized measure (13) remains to be clarified. It is still too early to tell what the most useful approach will be, but one may note that the emphasis on the curvature shares the basic weakness in the PPM approach in dealing with oscillations.

Other researchers have looked directly at the *dynamic decomposition weights (DDW)* w_i in (10), i. e., the relative weight of the modes for a particular variable, from a policy criterion perspective, similar to Forrester's original focus and the starting point for the EEA analysis [21,53,54].

For instance, Saleh et al. [54] look at how alternative stabilization policies affect the behavior of business cycle models, using both a simple inventory–workforce model [59], and a more extensive model based on Mass [32] and used in the LEEA analysis of Kampmann and Oliva [29]. Using the procedure in Fig. 9, they decompose the net stabilizing effect of a policy into its effect on the behavior mode itself (LEEA) and its effect on the expression of that mode in the variable of interest, measured the dynamic decomposition weights (EVA or DDW).

To illustrate the approach we perform the computations for the inventory–workforce model (Fig. 3 and Table 1). We find that the following equations describe the behavior of the state variables

$$\begin{aligned} \text{ED} &= 1 - 0.500\text{e}^{-0.250t} \\ \text{INV} &= 3 - 2.167\text{e}^{-0.250t} \\ &\quad + 1.134\text{e}^{-0.138t} \sin(2.945 + 0.285t) \\ \text{WF} &= 1 + 0.669\text{e}^{-0.250t} \\ &\quad - 1.169\text{e}^{-0.138t} \sin(1.553 - 0.285t)\ . \end{aligned} \tag{14}$$

As expected from the structure of the model, the behavior of *Expected Demand* does not have an oscillatory component and only shows a short transient exponential adjustment for the stock to match *Demand*. On the other hand, *Inventory* and *Workforce*, in addition to having the transient behavior to reach equilibrium captured by the first eigenvalue, have an oscillatory component represented by the second eigenvalue. Note that each state variable has a different *Dynamic Decomposition Weight* (w) for each

System Dynamics, Analytical Methods for Structural Dominance Analysis in, Table 4

Elasticity to parameters of weight of eigenvalue 2 (−0.138 + i0.85) on inventory and influence of parameters on eigenvalue 2 – inventory–workforce model

Parameter	w_2 on INV Elasticity	Influence on Re[$\boldsymbol{\lambda}_2$]	Influence on Im[$\boldsymbol{\lambda}_2$]
Demand	2.000	0.000	0.000
Inventory correction time	0.656	0.038	−0.157
Flexibility in production	0.364	−0.549	−0.266
Productivity	−0.353	0.000	0.000
Time to change expectations	−0.240	0.000	0.000
Normal inventory coverage	0.239	0.000	0.000
Hiring/Firing time	0.238	0.100	−0.127
Disequilibrium fraction	0.000	0.000	0.000

reference mode, i. e., each eigenvalue contributes differently to the overall behavior of each state variable.

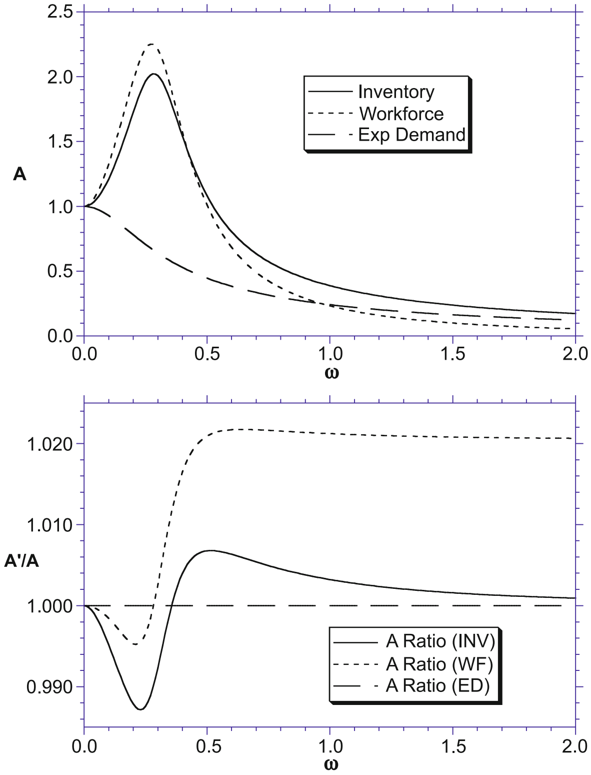
An exploration of the policy design space can be achieved by assessing the influence of model parameters on the dynamic decomposition weight. By focusing on the weights of the behavior modes for the variable of interest we can identify leverage points to increase or decrease the presence of a behavior mode in the variable. The weight elasticity column in Table 4 reports the parameter elasticity of w_2 (the weight of eigenvalue 2, the oscillatory behavior mode) on *Inventory* ($\varepsilon_w = (\mathrm{d}w/\mathrm{d}p)(p/w)$). The magnitude of the elasticity quantifies the impact that changes in the parameter value have on the weight of the oscillatory behavior model on *Inventory*. The table is sorted in descending order of absolute value of elasticity.

Changes in parameters, however, not only impact the behavior decomposition weights, but also change the eigenvalues themselves. This dual impact of parameter changes introduces a challenge in developing policy recommendations. The last two columns of Table 4 report the influence on the eigenvalue (real and imaginary part) for each parameter. These measures of influence should be interpreted in a similar way as the weight elasticities. The influence measure is defined as $\mu_\lambda = \left(\partial\lambda/\partial p\right) p$. A positive real-part measure indicates that increasing the parameter will destabilize the system by lengthening the settling time and vice-versa. A positive imaginary-part measure indicates that increasing the parameter will increase the frequency of oscillation – normally considered a destabilizing influence – and vice-versa.

Five parameters, *demand, disequilibrium fraction*, *productivity, normal inventory coverage*, and *time to change expectations*, have no influence on the oscillatory behavior mode. *Demand* and *disequilibrium fraction* are initialization constants that do not participate in any of the feedback loops in the model. *Productivity* is essentially a scaling measure having to do with the definition of units of labor and goods in the model. Redefining units should not affect the dynamics of the model. While *time to change expectations* is involved in loop 1, it does not participate in the oscillatory behavior observed in the model since, as discuss above, *Expected Demand* is in a separate strongly connected component of the model.

In accordance with LEEA, the *flexibility in production* parameter, which strengthens overtime loop 3 (cf. Fig. 3), has a strong stabilizing influence, by both increasing the damping and lowering the frequency of the oscillatory mode. Likewise, as predicted by LEEA, a shorter *hiring/firing time* will increase damping by strengthening the labor adjustment loop 5 but, again in accordance with LEEA, also increases the frequency of adjustment because it also strengthens the major loop 4. Finally, lowering the *inventory correction time* will strengthen the link from *inventory* to *desired production*, and consequently the three loops 3, 4 and 5, with the net effect that although the adjustment is a little faster (a more negative real part), the frequency is also increased significantly, i. e., it is a less effective way of stabilizing the system (cf. Fig. 9).

As an alternative approach, Fig. 11 shows what happens to the frequency response of the state variables (*Inventory* INV, *Workforce* WF, and *Expected Demand* ED) when the parameter Hiring/Firing Time (hft) is reduced by 2% from 5 to 4.9. There are a number of things to notice in the figure. First, there is no effect whatsoever on the ED variable, which should not be surprising, given that there is no feedback to this variable from the rest of the system. Second, the effect on the amplitude, like the amplitude itself, is strongly dependent upon the frequency of variation. We see that there is a significant amount of dampening on the *Inventory* fluctuation around the resonant frequencies in the range 0.1 to 0.3. On the other hand, there is a small amount of amplification of inventory in the higher frequency ranges. The effect on *Workforce* is very different: though there is a small attenuation in the reso-



System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 11

Effect on frequency response of the inventory workforce model of reducing the parameter Hiring/Firing Time (hft) from 5.0 to 4.9. The diagram shows the gain of the base case (*upper graph*) for the three state variables, and the resulting change in the gain, measured as the ratio (A'/A) from the parameter change (*lower graph*)

nant frequencies, there is a significant increase in variance in the higher frequency range. In other words, although the LEEA analysis showed a faster hiring policy to be stabilizing (by strengthening loop 2, cf. Table 3), the DDW analysis shows that it depends – both upon the variable in question and the context (frequency of variation).

Future Directions

As mentioned above, it is not possible to construct a complete theory that will automatically provide modelers with "the" dominant structure. Given the analytical intractability of nonlinear high-order systems found in our field, the most we can hope for is a set of tools that will guide the analysis and aid the development of the modeler's intuition.

That said, however, we are left with an impression that the analytical foundation for these tools is in need of further development before one rushes into implementing them into software packages. We are quite satisfied with the current state of affairs in this regard, where code, models, and documentation are made freely to download (most of the cited papers provide a URL to their code and models). Understanding *how* and *why* the tools work the way they do is crucial, and this will require that a number of puzzles, uncertainties, and technical problems be addressed. Only then will the time come to submit the methods for wider application to test their real-world utility.

While the classical method remains a useful intuitive guide and teaching tool for graduate students, there are no signs that it may be developed further. (That said, it is possible that the classical control transfer function method may be employed in the eigensystem approaches to explore nested canonical systems, though this is purely speculative). The pathway method would benefit from a firmer mathematical foundation. In particular, it would be important to compare how its results and conclusions compare to those found in the LST. It is possible that the pathway method may eventually be merged with the LST approaches as a subset of a general analytical toolbox. We believe that there is a great deal of promise in combining the eigenvalue and eigenvector analysis in the LST approaches. This combination will yield a complete system characterization and an understanding of both how particular feedback loops are involved in generating a behavior mode, and how system elements determine the expression of that behavior mode in a particular variable. A unified LST approach along the lines suggested in Fig. 10 thus seems within reach.

It will probably be a while, however, before these methods will find their way into widely available and use-friendly software packages. Apart from the theoretical issues alluded to above, a number of technical issues related to numerical calculations, various "pathological cases" (such as non-distinct eigenvalues), and special cases of feedback loops ("figure-eight" loops, for instance), will need to be addressed.

On the more creative side, it would be interesting to explore alternative forms of visualizing the various influence measures developed. For instance, one could imagine that links between variables in a model diagram "glow" in different colors and intensities depending upon their effect on a behavior pattern in question. This is not just a question of fancy user interfaces: as mentioned in the introduction, the function of these tools will be as intuitive consistent aids to understanding, not analytical "answering machines". In this light, the visualization is as important as the analytical principles behind it. Given the power of the human eye in finding patterns in visual data, this could be a significant next step.

Bibliography

1. Abdel-Gawad A, Abdel-Aleem B, Saleh M, Davidsen P (2005) Identifying dominant behavior patterns, links and loops: Automated eigenvalue analysis of system dynamics models. Proceedings of the Int System Dynamics Conference, Boston, July 2005. System Dynamics Society, Albany
2. Barlas Y (1989) Multiple Tests for Validation of System Dynamics Type of Simulation Models. Eur J Oper Res 42(1):59–87
3. Bass FM (1969) A new product growth for model consumer durables. Manag Sci 15(5):215–227
4. Chen CT (1970) Introduction to Linear System Theory. Holt, Rinnehart and Winston, New York
5. Eberlein R (1984) Simplifying Models by Retaining Selected Behavior Modes. Ph D Thesis, Sloan School of Management, MIT, Cambridge
6. Eberlein RL (1986) Full Feedback Parameter Estimation. In: Proceedings of the Int Systems Dynamics Conference, Sevilla, Spain. System Dynamics Society, Albany, pp 69–83
7. Eberlein RL (1989) Simplification and understanding of models. Syst Dyn Rev 5(1):51–68
8. Eberlein RL, Wang Q (1985) Statistical Estimation and System Dynamics Models. Proceedings of the Int Systems Dynamics Conference, Keystone, USA. System Dynamics Society, Albany, pp 206–222
9. Ford A, Flynn H (2005) Statistical screening of system dynamics models. Syst Dyn Rev 21(4):273–303
10. Ford DN (1999) A Behavioral Approach to Feedback Loop Dominance Analysis. Syst Dyn Rev 15(1):3–36
11. Forrester JW (1961) Industrial Dynamics. Productivity Press, Cambridge
12. Forrester JW (1969) Urban Dynamics. Productivity Press, Cambridge
13. Forrester JW (1971) World Dynamics. Productivity Press, Cambridge
14. Forrester N (1982) A Dynamic Synthesis of Basic Macroeconomic Policy: Implications for Stabilization Policy Analysis. Ph D Thesis, Sloan School of Management, MIT, Cambridge
15. Forrester N (1983) Eigenvalue analysis of dominant feedback loops. Proceedings of the Int System Dynamics Conference, Chestnut Hill, USA. System Dynamics Society, Albany
16. Forrester JW (1993) System Dynamics and the Lessons of 35 Years. In: DeGreene KB (ed) Systems-Based Approach to Policymaking. Kluwer, Norwell, pp 199–240
17. Forrester JW, Senge PM (1980) Tests for Building Confidence in System Dynamics Models. TIMS Stud Manag Sci 14:209–228
18. Forrester JW, Mass NJ, Ryan CJ (1976) The System Dynamics National Model: Understanding Socio-Economic Behavior and Policy Alternatives. Technol Forecast Soc Chang 9(1–2):51–68
19. Goldberg DE (1989) Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Reading
20. Gonçalves P (2003) Demand bubbles and phantom orders in supply chains. Ph D Thesis, Sloan School of Management, MIT, Cambridge
21. Gonçalves P (2008) Behavior modes, pathways and overall trajectories: Eigenvalue and eigenvector analysis in system dynamics. Syst Dyn Rev, forthcoming
22. Gonçalves P, Lerpattarapong C, Hines JH (2000) Implementing formal model analysis. Proceedings of the Int System Dynamics Conference, Bergen, Norway, August 2000. System Dynamics Society, Albany
23. Graham AK (1977) Principles on the Relationship Between Structure and Behavior of Dynamic Systems. Ph D Thesis, Sloan School of Management, MIT, Cambridge
24. Güneralp B (2006) Towards Coherent Loop Dominance Analysis: Progress in Eigenvalue Elasticity Analysis. Syst Dyn Rev 22(3):263–289
25. Haxholdt C, Kampmann CE, Mosekilde E, Sterman JD (1995) Mode Locking and Entrainment of Endogenous Economic Cycles. Syst Dyn Rev 11(3):177–198
26. Holland JH (1992) Adaptation in Natural and Artificial Systems. MIT Press, Cambridge
27. Homer JB (1983) Partial-Model Testing as a Validation Tool for System Dynamics. In: Proceedings of the Int System Dynamics Conference, Chestnut Hill, USA, July 1983. System Dynamics Society, Albany, pp 920–932
28. Kampmann CE (1996) Feedback Loop Gains and System Behavior (unpublished manuscript). In: Proceedings of the Int System Dynamics Conference, Cambridge, USA, July 1996. System Dynamics Society, Albany, pp 260–263
29. Kampmann CE, Oliva R (2006) Loop Eigenvalue Elasticity Analysis: Three Case Studies. Syst Dyn Rev 22(2):146–162
30. Lane DC, Smart C (1996) Reinterpreting 'generic structure': evolution, application and limitations of a concept. Syst Dyn Rev 12(2):87–120
31. Luenberger DG (1979) Introduction to Dynamic Systems: Theory, Models and Applications. Wiley, New York
32. Mass NJ (1975) Economic Cycles: An Analysis of Underlying Causes. Productivity Press, Cambridge
33. Meadows DH, Meadows DL, Randers J, Behrens III WW (1972) The Limits to Growth: A Report for the Club of Rome's Project on the Predicament of Mankind. Universe Books, New York
34. Mojtahedzadeh MT (1996) A path taken: Computer-assisted heuristics for understanding dynamic systems. Ph D Thesis, Rockefeller College of Pubic Affairs and Policy, State University of New York at Albany, Albany
35. Mojtahedzadeh MT, Andersen D, Richardson GP (2004) Using Digest to implement the pathway participation method for detecting influential system structure. Syst Dyn Rev 20(1):1–20
36. Morecroft JDW (1985) Rationality in the Analysis of Behavioral Simulation Models. Manag Sci 31(7):900–916
37. Mosekilde E, Laugesen JL (2006) Nonlinear Dynamic Phenomena in the BEER Model. Department of Physics, The Technical University of Denmark, Kongens Lyngby
38. Ogata K (1990) Modern Control Engineering, 2nd edn. Prentice Hall, Englewood Cliffs
39. Oliva R (2003) Model Calibration as a Testing Strategy for System Dynamics Models. Eur J Operat Res 151(3):552–568
40. Oliva R (2004) Model Structure Analysis Through Graph Theory: Partition Heuristics and Feedback Structure Decomposition. Syst Dyn Rev 20(4):313–336
41. Oliva R, Mojtahedzadeh M (2004) Keep it simple: Dominance assessment of short feedback loops. Proceedings of the Int System Dynamics Conference, Oxford, UK, July 2004. System Dynamics Society, Albany
42. Ott E (1993) Chaos in Dynamical Systems. Cambridge University Press, New York

43. Peterson DW (1980) Statistical Tools for System Dynamics. In: Randers J (ed) Elements of the System Dynamics Method. Productivity Press, Cambridge, pp 224–241
44. Peterson DW, Eberlein RL (1994) Reality Checks: A Bridge Between Systems Thinking and System Dynamics. Syst Dyn Rev 10(2/3):159–174
45. Radzicki MJ (2004) Expectation Formation and Parameter Estimation in Uncertain Dynamical Systems: The System Dynamics Approach to Post Keynesian-Institutional Economics. Proceedings of the Int System Dynamics Conference, Oxford, UK, July 2004. System Dynamics Society, Albany
46. Richardson GP (1984/1995) Loop Polarity, Loop Dominance, and the Concept of Dominant Polarity. Syst Dyn Rev 11(1):67–88
47. Richardson GP (1986) Dominant structure. Syst Dyn Rev 2(1):68–75
48. Richardson GP (ed) (1988) System Dynamics Review. Chaos Special Issue 4:1–2
49. Richmond B (1980) A new look at an old friend. Plexus, Resource Policy Center, Thayer School of Engineering, Dartmouth College, Hanover
50. Saleh M (2002) The characterization of model behavior and its causal foundation. Ph D Thesis, Dept of Information Science, University of Bergen, Bergen
51. Saleh M, Davidsen P (2001) The origins of behavior patterns. Proceedings of the Int System Dynamics Conference, Atlanta, July 2001. System Dynamics Society, Albany
52. Saleh M, Davidsen P (2001) The origins of business cycles. Proceedings of the Int System Dynamics Conference, Atlanta, July 2001. System Dynamics Society, Albany
53. Saleh M, Oliva R, Davidsen P, Kampmann CE (2006) Eigenvalue Analysis of System Dynamics Models: Another Perspective. Proceedings of the Int System Dynamics Conference, Neijmegen, The Netherlands, July 2006. System Dynamics Society, Albany
54. Saleh M, Oliva R, Davidsen P, Kampmann CE (2008) A comprehensive analytical approach for policy analysis of system dynamics models. Mays Business School, Texas A&M University, Working Paper
55. Schweppe F (1973) Uncertain Dynamical Systems. Prentice-Hall, Englewood Cliffs
56. Senge PM (1990) The Fifth Discipline: The Art & Practice of the Learning Organization. Doubleday Currency, New York
57. Sterman JD (1981) The Energy Transition and the Economy: A System Dynamics Approach. Ph D Thesis, Sloan School of Management, MIT, Cambridge
58. Sterman JD (1985) A Behavioral Model of the Economic Long Wave. J Econ Behav Org 6(1):17–53
59. Sterman JD (2000) Business dynamics: Systems thinking and modeling for a complex world. Irwin McGraw-Hill, Boston
60. Wolstenholme E (2004) Using generic system archetypes to support thinking and modelling. Syst Dyn Rev 20(4):341–356
61. Zhusubaliyev ZT, Mosekilde E (2003) Bifurcation and Chaos in Piecewise-Smooth Dynamical Systems. World Scientific, Singapore

System Dynamics, The Basic Elements of

GEORGE P. RICHARDSON
Rockefeller College of Public Affairs and Policy,
University at Albany, State University of New York,
Albany, USA

Article Outline

Glossary

Endogenous Generated from within. Contrasting with "exogenous," meaning generated by forces external to a system or point of view.

Feedback loop A closed path of causal influences and information, forming a circular-causal loop of information and action.

System dynamics System dynamics is a computer-aided approach to theory-building, policy analysis and strategic decision support emerging from an endogenous point of view.

Definition of the Subject

System dynamics is a computer-aided approach to theory-building, policy analysis, and strategic decision support emerging from an endogenous point of view [18,20]. It applies to dynamic problems arising in complex social, managerial, economic, or ecological systems – literally any dynamic systems characterized by interdependence, mutual interaction, information feedback, and circular causality.

Introduction

The field of system dynamics developed initially from the work of Jay W. Forrester. His seminal book *Industrial Dynamics* [7] is still a significant statement of philosophy and methodology in the field. Within ten years of its publication, the span of applications grew from corporate and industrial problems to include the management of research and development, urban stagnation and decay, commodity cycles, and the dynamics of growth in a finite world. It is now applied in economics, public policy, environmental studies, defense, theory-building in social science, and other areas, as well as its home field, management. The name industrial dynamics no longer does justice to the breadth of the field (for extensive examples, see [20,28], so it has become generalized to system dynamics. The modern name suggests links to other systems methodologies, but the links are weak and misleading. System dynamics emerges out of servomechanisms engineering, not general systems theory or cybernetics [18].

The system dynamics approach involves:

- Defining problems dynamically, in terms of graphs over time.
- Striving for an endogenous, behavioral view of the significant dynamics of a system, a focus inward on the characteristics of a system that themselves generate or exacerbate the perceived problem.
- Thinking of all concepts in the real system as continuous quantities interconnected in loops of information feedback and circular causality.
- Identifying independent stocks or accumulations (levels) in the system and their inflows and outflows (rates).
- Formulating a behavioral model capable of reproducing, by itself, the dynamic problem of concern. The model is usually a computer simulation model expressed in nonlinear equations, but is occasionally left unquantified as a diagram capturing the stock-and-flow/causal feedback structure of the system.
- Deriving understandings and applicable policy insights from the resulting model.
- Implementing changes resulting from model-based understandings and insights.

Mathematically, the basic structure of a formal system dynamics computer simulation model is a system of coupled, nonlinear, first-order differential (or integral) equations,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}(t) = \mathbf{f}(\mathbf{x}, \mathbf{p}) \, ,$$

where $\mathbf{x}$ is a vector of levels (stocks or state variables), $\mathbf{p}$ is a set of parameters, and $\mathbf{f}$ is a nonlinear vector-valued function. Such a system has been variously called a *state-determined system* in the engineering literature, an *absolute system* [3], an *equifinal system* [32], and a *dynamical system* [16].

Simulation of such systems is easily accomplished by partitioning simulated time into discrete intervals of length $\mathrm{d}t$ and stepping the system through time one $\mathrm{d}t$ at a time. Each state variable is computed from its previous value and its net rate of change $x'(t)$: $x(t) = x(t - \mathrm{d}t) + \mathrm{d}t \cdot x'(t - \mathrm{d}t)$. In the earliest simulation language in the field (DYNAMO) this equation was written with time scripts K (the current moment), J (the previous moment), and JK (the interval between time J and K): $X_{\mathrm{K}} = X_{\mathrm{J}} + \mathrm{DT} \cdot \mathrm{XRATE}_{\mathrm{JK}}$ (see, e. g., [22]). The computation interval $\mathrm{d}t$ is selected small enough to have no discernible effect on the patterns of dynamic behavior exhibited by the model. In more recent simulation environments, more sophisticated integration schemes are available (although the equation written by the user may look like this simple Euler integration scheme), and time scripts may not be in evidence. Important current simulation environments include STELLA and iThink (isee Systems, http://www.iseesystems.com/), Vensim (Ventana Systems, http://www.vensim.com/), and Powersim (http://www.powersim.com/).

Forrester's original work stressed a continuous approach, but increasingly modern applications of system dynamics contain a mix of discrete difference equations and continuous differential or integral equations. Some practitioners associated with the field of system dynamics work on the mathematics of such structures, including the theory and mechanics of computer simulation, analysis and simplification of dynamic systems, policy optimization, dynamical systems theory, and complex nonlinear dynamics and deterministic chaos.

The main applied work in the field, however, focuses on understanding the dynamics of complex systems for the purpose of policy analysis and design. The conceptual tools and concepts of the field – including feedback thinking, stocks and flows, the concept of feedback loop dominance, and an endogenous point of view – are as important to the field as its simulation methods.

Feedback Thinking

Conceptually, the feedback concept is at the heart of the system dynamics approach. Diagrams of loops of information feedback and circular causality are tools for conceptualizing the structure of a complex system and for communicating model-based insights. Intuitively, a feedback loops exists when information resulting from some action travels through a system and eventually returns in some form to its point of origin, potentially influencing future action. If the tendency in the loop is to reinforce the initial action, the loop is called a *positive* or *reinforcing* feedback loop; if the tendency is to oppose the initial action, the loop is called a *negative, counteracting*, or *balancing* feedback loop. The sign of the loop is called its *po-*

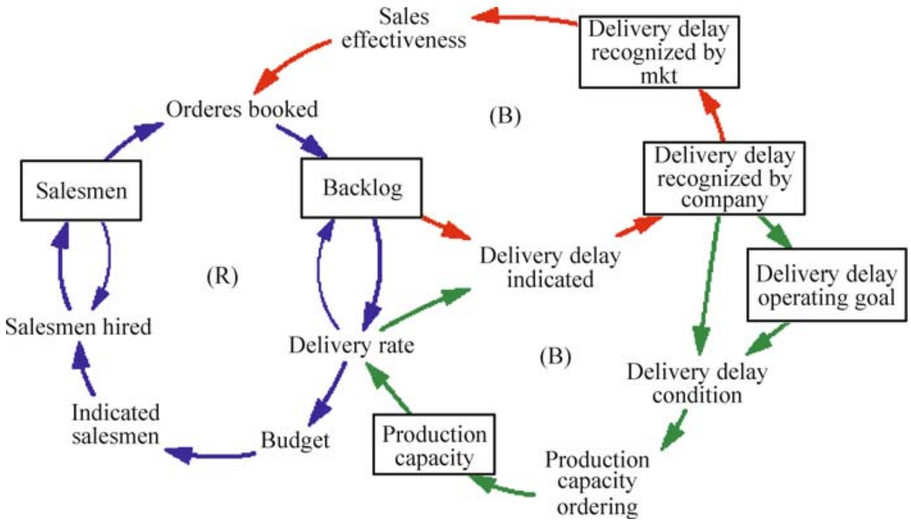
larity. Balancing loops can be variously characterized as goal-seeking, equilibrating, or stabilizing processes. They can sometimes generate oscillations, as when a pendulum seeking its equilibrium goal gathers momentum and overshoots it. Reinforcing loops are sources of growth or accelerating collapse; they are disequilibrating and destabilizing. Combined, balancing and reinforcing circular causal feedback loops can generate all manner of dynamic patterns.

Feedback loops are ubiquitous in human and natural systems and, under various names and representations, have been widely recognized in popular and scholarly literature. Feedback thought has been present implicitly or explicitly for hundreds of years in the social sciences and literally thousands of years in recorded history [9]. We have the vicious circle originating in classical logic and morphing into common usage, the bandwagon effect, the invisible hand of Adam Smith, Malthus's correct observation of population growth as a self-reinforcing process, Keynes's consumption multiplier, the investment accelerator of Hicks and Samuelson, compound interest or inflation, the biological concepts of proprioception and homeostasis, Festinger's cognitive dissonance, Myrdal's principle of cumulative causation, Venn's idea of a suicidal prophecy, Merton's related notion of a self-fulfilling prophecy, and so on. Each of these ideas can be concisely and insightfully represented as one or more loops of causal influences with positive or negative polarities. Great social scientists and feedback thinkers; great social theories are feedback thoughts. (For a full exposition of the evolution of the feedback concept see [19].)

Loop Dominance and Nonlinearity

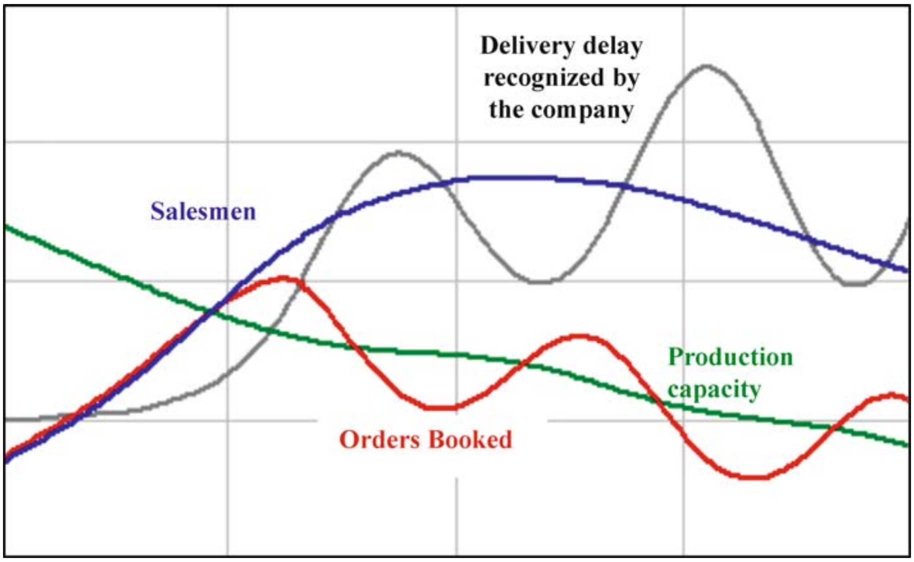
The loop concept underlying feedback and circular causality by itself is not enough, however. The explanatory power and insightfulness of feedback understandings also rest on the notions of active structure and loop dominance. Complex systems change over time. A crucial requirement for a powerful view of a dynamic system is the ability of a mental or formal model to change the strengths of influences as conditions change, that is to say, the ability to shift *active* or *dominant structure*.

In a system of equations, this ability to shift loop dominance comes about endogenously from nonlinearities in the system. For example, the S-shaped dynamic behavior of the classic logistic growth model ($\mathrm{d}P/\mathrm{d}t = aP - bP^2$) or similar structures like the Gompertz curve ($\mathrm{d}P/\mathrm{d}t = aP - bP\ln(P)$) can be seen as the consequence of a shift in loop dominance from a positive, self-reinforcing feedback loop (aP) producing exponential-like growth, to a negative feedback loop ($-bP^2$ or $-bP\ln(P)$) that brings the system to its eventual goal. The shift in loop dominance in these models comes about from the nonlinearity in the second term, which grows faster than the first term and eventually overtakes it. Only nonlinear models can endogenously alter their active or dominant structure and shift loop dominance.



System Dynamics, The Basic Elements of, Figure 1

Core structure of Forrester's market growth model [8], showing a *blue reinforcing loop* underlying the growth (or reinforcing decline) of Salesmen, Orders, and Revenue, a *red balancing loop* containing various delayed recognitions of the company's delivery delay, and a *green balancing loop* responsible for capacity ordering if the delivery delay drops too far below its operating goal



System Dynamics, The Basic Elements of, Figure 2
The dynamic behavior of the model shown in Fig. 1, illustrating an early growth phase, which turns into an oscillatory phase as the feedback loop dominance shifts to the *red balancing delivery delay loop*, and results in a long term corporate decline as the *green capacity ordering loop* responds to a sliding operating goal for the acceptable delivery delay

Real systems are perceived to change their active or dominant structure over time, often because of the build-up of internal forces. Thus from a feedback perspective, the ability of nonlinearities to generate shifts in loop dominance is the fundamental reason for advocating nonlinear models of social system behavior.

Figures 1 and 2, abstracted from an early, classic paper [8] illustrate these ideas. In Fig. 1 salesmen (in the blue reinforcing loop) book orders for the company; if enough revenue is generated, there is enough budget to hire more salesmen and corporate growth ensues. Whether salesmen (in this simplified picture) book enough orders depends on the company's delivery delay for the product, as perceived by the market (red balancing loop). The company builds production capacity according to its perceived need, as indicated by its perceived delivery delay and its target for that (green balancing loop).

Figure 2 shows the dynamics this feedback structure endogenously generates. In the early phase, salesmen grow as orders and revenue grow; the system's exponential growth behavior in that phase is generated by the reinforcing salesmen loop. But then the feedback loop dominance soon shifts to the balancing delivery delay loop, which constrains sales effectiveness and brings a halt to growth. The system moves into an oscillatory phase generated by the various monitoring and perception delays around the now dominant red balancing loop. Salesmen eventual peak and decline, as the green production capacity ordering loop fails to keep production capacity sufficient to hold the delivery delays in check.

Thus the dynamic behavior of this system is a consequence of its feedback structure and the nonlinearities that shift loop dominance endogenously over time. The particular decline scenario shown in Fig. 2 illustrates one of the deep insights of the model: the adaptive goal structure, in which the delivery delay operating goal moves slowly to accommodate changes in the company's delivery delay, weakens the green balancing loop trying to bring on capacity. The company never perceives its delivery delay is sufficiently higher than its (sliding) target, so it fails to order sufficient capacity to sustain growth. A fixed goal for the acceptable delivery delay sends a stronger signal, which can turn this corporate decline into oscillating growth [8].

Thus, nonlinearity is crucial to the system dynamics approach. However, it is crucial not merely because of its mathematical properties but because it enables the formalization of a profoundly powerful perspective on theory and policy – the *endogenous point of view*.

The Endogenous Point of View

The concept of endogenous change is fundamental to the system dynamics approach. It has both philosophical and engineering origins. A deep and lasting insight of the earliest attempts at servomechanisms control is the realization that *the attempt to control a system generates dynamics of* *its own*, complicating the dynamics trying to be controlled. A governor mechanism imposed to control the speed of a steam engine can generate oscillatory "hunting behavior," as the control system overshoots and undershoots the set point. As it becomes part of the system, the governing mechanism thus generates dynamics of its own.

The insight transfers readily, but with added significance, from engineering systems to people systems: Attempts to control complex human systems – coercing, guiding, managing, governing – generate dynamics of their own. Moreover, some of these endogenously generated dynamics are created by the control mechanisms themselves (like the governor of a steam engine) and some are created by human creative responses to the management efforts (e. g., principal-agent interactions). These natural and human forces, creating counteracting and compensating pressures in response to system control efforts, emerge as complicated circular-causal feedback structures. The often complex, difficult-to-understand dynamics of such management systems are to a great degree a consequence of their internal structures.

To capture and analyze such management complexities, one must look inward to see the ways a complex system naturally responds to system pressures. The endogenous point of view is thus central to the system dynamics approach. It dictates aspects of model formulation: exogenous disturbances are seen at most as *triggers* of system behavior (like displacing a pendulum); the *causes* are contained within the structure of the system itself (like the interaction of a pendulum's position and momentum that produces oscillations). Corrective responses are also not modeled as functions of time, but are dependent on conditions within the system. Time by itself is not seen as a cause in the endogenous point of view.

Theory building and policy analysis are significantly affected by this endogenous perspective. Taking an endogenous view exposes the natural *compensating* tendencies in social systems that conspire to defeat many policy initiatives. Feedback and circular causality are delayed, devious, and deceptive. For understanding, system dynamics practitioners strive for an *endogenous point of view*. The effort is to uncover the sources of system behavior that exist within the structure of the system itself.

System Structure

These ideas are captured almost explicitly in Forrester's [9] organizing framework for system structure:

- Closed boundary
- Feedback loops
- Levels
- Rates
- Goal
- Observed condition
- Discrepancy
- Desired action.

The *closed boundary* signals the endogenous point of view. The word *closed* here does not refer to open and closed systems in the general system sense, but rather refers to the effort to view a system as *causally* closed. The modeler's goal is to assemble a formal structure that can, *by itself*, without exogenous explanations, reproduce the essential characteristics of a dynamic problem.

The causally closed system boundary at the head of this organizing framework identifies the endogenous point of view as the feedback view pressed to an extreme. Feedback thinking can be seen as a *consequence* of the effort to capture dynamics within a closed causal boundary. Without causal loops, all variables must trace the sources of their variation ultimately outside a system. Assuming instead that the causes of all significant behavior in the system are contained within some closed causal boundary forces causal influences to feed back upon themselves, forming causal loops. Feedback loops enable the endogenous point of view and give it structure.

Levels and Rates

Stocks (accumulations, or "levels" in early system dynamics literature) and the flows ("rates") that affect them are essential components of system structure. A map of causal influences and feedback loops is not enough to determine the dynamic behavior of a system. A constant inflow yields a linearly rising stock; a linearly rising inflow yields a stock rising along a parabolic path; a stock with inflow proportional to itself grows exponentially; two stocks in a balancing loop have a tendency to generate oscillations; and so on. For example, the boxes in Fig. 1 represent accumulations in the company and its market; the three stocks in the red balancing loop (the order backlog and the two perceptions of the company's delivery delay) give that loop its tendency to generate oscillations which propagate through out the system. Accumulations are the memory of a dynamic system and contribute to its disequilibrium and dynamic behavior.

Forrester [7] placed the operating policies of a system among its rates, the inflows and outflows governing change in the system. Many of these rates of change assume the classic structure of a negative feedback loop striving to take action to reduce the discrepancy between the observed condition of the system and a goal. The simplest such rate structure results in an equation of the form

$$\text{RATE} = \frac{\text{GOAL} - \text{LEVEL}}{\text{ADJUSTMENT TIME}} ,$$

where ADJUSTMENT TIME is the time over which the level adjusts to reach the goal. This simple formulation reflects Forrester's more general statement about rates in his hierarchy of system structure (above) which can be richly thought of as

$$\text{RATE} = f(\text{DESIRED ACTION})$$

$$\begin{aligned}&\text{DESIRED ACTION}\\ &= g(\text{DESIRED CONDITION},\\ &\qquad\qquad\qquad \text{OBSERVED CONDITION})\end{aligned}$$

$$\text{OBSERVED CONDITION} = h(\text{LEVELS}) ,$$

for some functions f, g, and h representing particular system characteristics.

Operating policies in a management system can influence the *flows* of information, material, and resources, which are the only means of changing the accumulations in the system. While flows can be changed quickly, as a matter of relatively quick decision making, stocks change slowly – they rise when inflows are great than outflows, and decline when inflows are less than outflows.

The simple "tub dynamics" of stocks are clear even to children, yet can be befuddling in complex systems. The accumulation of green house gases in the atmosphere, for example, affects the *flow* of heat energy radiated from the earth. To turn around global warming, the *accumulation* of green house gases must drop far enough to raise radiant energy above the inflow of solar energy, a simple stock-and-flow insight. But to cause the accumulation of green house gases to drop, their generation must fall below their natural absorption rate (another simple stock-and-flow observation). So turning around global warming is a process involving a chain of at least two significant accumulations, and people have trouble thinking it through reliably. The accumulations can only be changed by managing their associated flows. They will change only slowly even if we manage the technical and political pitfalls involved in lowering green house gas production (see [29]).

The significance of stocks in complex systems is vivid in a resource-based view of strategy and policy. Resources that enable a corporation or government to function or flourish are stocks, usually accumulated over long periods of time with significant investment of time, energy, and money. Reputations are also stocks, built over similarly long periods of time. While inadequate by themselves to give a full picture of the dynamics of a complex system, stocks and flows are vital components of system structure, without which fundamental understandings of dynamics are impossible [33].

Behavior is a Consequence of System Structure

The importance of stocks and flows appears most clearly when one takes a *continuous* view of structure and dynamics. Although a discrete view, focusing on separate events and decisions, is entirely compatible with an endogenous feedback perspective, the system dynamics approach emphasizes a continuous view [7]. The continuous view strives to look beyond events to see the dynamic patterns underlying them: model not the appearance of a discrete new housing unit in a city, but focus instead on the rise and fall of aggregate numbers of housing units. Moreover, the continuous view focuses not on discrete decisions but on the *policy structure* underlying decisions: not why this particular apartment building was constructed but what persistent pressures exist in the urban system that produce decisions that change housing availability in the city. Events and decisions are seen as surface phenomena that ride on an underlying tide of system structure and behavior. It is that underlying tide of policy structure and continuous behavior that is the system dynamicist's focus.

There is thus a *distancing* inherent in the system dynamics approach – not so close as to be confused by discrete decisions and myriad operational details, but not so far away as to miss the critical elements of policy structure and behavior. Events are deliberately blurred into dynamic behavior. Decisions are deliberately blurred into perceived policy structures. Insights into the connections between system structure and dynamic behavior, which are the goal of the system dynamics approach, come from this particular distance of perspective.

Suggestions for Further Reading on the Core of System Dynamics

The *System Dynamics Review*, the journal of the System Dynamics Society, published by Wiley, is the best source of current activity in the field, including methodological advances and applications.

The core of a vibrant field is difficult to discern in the flow of current work. However, the works that the field itself singles out as exemplary can give some reliable hints about what is considered vital to the core. In this sense two edited volumes are noteworthy: An early, interesting collection of applications is Roberts [24]; Richardson [21] is a more recent two-volume edited collection in the same spirit, containing prize-winning work in philosophical background, dynamic decision making, applications in the private and public sectors, and techniques for modeling with management.

In addition, the following works, selected from among winners of the System Dynamics Society's *Jay Wright Forrester Award* (see www.systemdynamics.org/Society_Awards.htm), can be considered insightful although implicit exemplars of the core of system dynamics. (Publications are listed beginning with the most recent; see the bibliography for full citations):

- Thomas S. Fiddaman, "Exploring policy options with a behavioral climate-economy model"
- Kim D. Warren, *Competitive Strategy Dynamics*
- Eric F. Wolstenholme, "Towards the Definition and Use of a Core Set of Archetypal Structures in System Dynamics"
- Nelson P. Repenning, "Understanding Fire Fighting in New Product Development"
- John D. Sterman, *Business Dynamics, Systems Thinking and Modeling for a Complex World*
- Peter Milling, "Modeling innovation processes for decision support and management simulation."
- Erling Moxnes, "Not Only the Tragedy of the Commons: Misperceptions of Bioeconomics."
- Jac A. M. Vennix, *Group Model Building: Facilitating Team Learning Using System Dynamics*
- Jack B. Homer, "A System Dynamics Model of National Cocaine Prevalence."
- Andrew Ford, "Estimating the Impact of Efficiency Standards on Uncertainty of the Northwest Electric System."
- Khalid Saeed, *Towards Sustainable Development: Essays on System Analysis of National Policy*
- Tarek Abdul-Hamid and Stuart Madnick, *Software Project Dynamics: An Integrated Approach*
- George P. Richardson, *Feedback Thought in Social Science and Systems Theory*
- Peter M. Senge, *The Fifth Discipline*
- John D. W. Morecroft, "Rationality in the Analysis of Behavioral Simulation Models."
- John D. Sterman, "Modeling Managerial Behavior: Misperceptions of Feedback in a Dynamic Decision Making Experiment."

For texts on the system dynamics approach, see Alfeld and Graham [2], Richardson and Pugh [22], Wolstenholme [34], Ford [6], Maani and Cavana [11], and the most comprehensive text to date, Sterman [28].

Bibliography

1. Abdul-Hamid T, Madnick S (1991) Software project dynamics: an integrated approach. Prentice Hall, Englewood Cliffs
2. Alfeld LE, Graham AK (1976) Introduction to urban dynamics. Pegasus Comunications, Waltham
3. Ashby H (1956) An Introduction to cybernetics. Chapman & Hall, London
4. Fiddaman TS (2002) Exploring policy options with a behavioral climate-economy model. Syst Dyn Rev 18(2):243–267
5. Ford A (1990) Estimating the impact of efficiency standards on uncertainty of the northwest electric system. Oper Res 38(4):580–597
6. Ford A (1999) Modeling the environment: an introduction to system dynamics of environmental systems. Island Press, Washington
7. Forrester JW (1961) Industrial dynamics. MIT Press, Cambridge. Reprinted by Pegasus Communications, Waltham
8. Forrester JW (1968) Market growth as influenced by capital investment. Ind Manag Rev (MIT, now Sloan Manag Rev) 9(2):83–105. Reprinted widely, e. g., Richardson [21]
9. Forrester JW (1969) Urban dynamics. MIT Press, Cambridge. Reprinted by Pegasus Communications, Waltham
10. Homer JB (1992) A system dynamics model of national cocaine prevalence. Syst Dyn Rev 9(1):49–78
11. Maani KE, Cavana RY (2000) Systems thinking and modelling: understanding change and complexity. Pearson Education, New Zealand
12. Milling P (1996) Modeling innovation processes for decision support and management simulation. Syst Dyn Rev 12(3):211–234
13. John DW, Morecroft JDW (1985) Rationality in the analysis of behavioral simulation models. Manag Sci 31(7):900–916
14. Morecroft JDW, Sterman JD (eds) (1994) Modeling for learning organizations. System dynamics series. Pegasus Communications, Waltham
15. Moxnes E (1998) Not only the tragedy of the commons: misperceptions of bioeconomics. Manag Sci 44(9):1234–1248
16. Nicholis G, Prigogine I (1977) Self-organization in nonequilibrium systems: from dissipative structures to order through fluctuations. Wiley, New York
17. Repenning NR (2001) Understanding fire fighting in new product development. J Prod Innov Manag 18(5):285–300
18. Richardson GP (1991) System dynamics: simulation for policy analysis from a feedback perspective. In: Fishwick PA, Luker PA (eds) Qualitative simulation modeling and analysis. Springer, New York
19. Richardson GP (1991) Feedback thought in social science and systems theory. University of Pennsylvania Press, Philadelphia. Reprinted by Pegasus Communications, 1999
20. Richardson GP (1996) System dynamics. In: Gass S, Harris C (eds) The encyclopedia of operations research and management science. Kluwer, New York
21. Richardson GP (ed) (1996) Modelling for management: simulation in support of systems thinking. International library of management. Dartmouth, Aldershot
22. Richardson GP, Pugh AL III. (1981) Introduction to system dynamics modeling with DYNAMO. MIT Press, Cambridge. Reprinted by Pegasus Communications, Waltham
23. Richmond B (1993) Systems thinking: critical thinking skills for the 1990s and beyond. Syst Dyn Rev 9:(2)113–133

24. Roberts EB (ed) (1978) Managerial applications of system dynamics. MIT Press, Cambridge. Reprinted by Pegasus Communications, Waltham
25. Saeed K (1991) Towards sustainable development: essays on system analysis of national policy. Progressive Publishers, Lahore
26. Senge PM (1990) The fifth discipline: the art and practice of the learning organization. Doubleday/Currency, New York
27. Sterman JD (1988) Modeling managerial behavior: misperceptions of feedback in a dynamic decision making experiment. Manag Sci 35(3):321–339
28. Sterman JD (2001) Business dynamics: systems thinking and modeling for a complex world. Irwin McGraw-Hill, Boston
29. Sterman JD, Sweeney LB (2002) Cloudy skies: assessing public understanding of global warming. Syst Dyn Rev 18(2):207–240
30. System Dynamics Review. 1985–present. Wiley, Chichester
31. Vennix JAM (1996) Group model building: facilitating team learning using system dynamics. Wiley, Chichester
32. von Bertalanffy L (1968) General systems theory: foundations, development, applications. George Braziller, New York
33. Warren KD (2002) Competitive strategy dynamics. Wiley, United Kingdom
34. Wolstenholme EF (1990) System enquiry: a system dynamics approach. Wiley, Chichester
35. Wolstenholme EF (2003) Towards the definition and use of a core set of archetypal structures in system dynamics. Syst Dyn Rev 19(1):7–26

System Dynamics in the Evolution of the Systems Approach

Markus Schwaninger
Institute of Management, University of St. Gallen,
St. Gallen, Switzerland

Article Outline

Glossary

Cybernetics The science of communication and control in complex, dynamical systems. The core objects of study are information, communication, feedback and adaptation. In the newer versions of cybernetics, the emphasis is on observation, self-organization, self-reference and learning.

Dynamical system The dynamical system concept is a mathematical formalization of time-dependent processes. Examples include the mathematical models that describe the swinging of a clock pendulum, the flow of water in a river, and the evolution of a population of fish in a lake.

Law of requisite variety Ashby's law of requisite variety says: "Only variety can destroy variety". It implies that the varieties of two interacting systems must be in balance, if stability is to be achieved.

Organizational cybernetics The science which applies cybernetic principles to organization. Synonyms are *Management Cybernetics* and *Managerial Cybernetics*.

System There are many definitions of *system*. Two examples: A portion of the world sufficiently well defined to be the subject of study; something characterized by a structure, for example, a social system (Anatol Rapoport). A system is a family of relationships between its members acting as a whole (International Society for the Systems Sciences).

System dynamics A methodology and discipline for the modeling, simulation and control of dynamic systems. The main emphasis falls on the role of structure and its relationship with the dynamic behavior of systems, which are modeled as networks of informationally closed feedback loops between stock and flow variables.

Systems approach A perspective of inquiry, education and management, which is based on system theory and cybernetics.

System theory A formal science of the structure, behavior, and development of systems. In fact there are different system theories. General system theory is a transdisciplinary framework for the description and analysis of any kind of system. System theories have been developed in many domains, e. g., mathematics, computer science, engineering, sociology, psychotherapy, biology and ecology.

Variety A technical term for *complexity* which denotes the number of (potential) states of a system.

Definition of the Subject

The purpose of this chapter is to give an overview of the role of system dynamics (SD) in the context of the evolution of the systems movement. This is necessary because SD is often erroneously taken as the systems approach as such, not as part of it. It is also requisite to show that the processes of the evolution of both SD in particular and the systems movement as a whole are intimately linked and intertwined. Finally, in view of the purpose of the chapter the actual and potential relationships between system dynamics and the other strands of the systems movement are evaluated. This way, complementarities and synergies are identified.

Introduction

The purpose of this contribution is to give an overview of the role of system dynamics in the context of the evolution of the systems movement. "Systems movement" – often referred to briefly as "systemics" – is a broad term, which takes into account the fact that there is no single system approach, but a range of different ones. The common denominator of the different system approaches in our day is that they share a worldview focused on complex dynamic systems, and an interest in describing, explaining and designing or at least influencing them. Therefore, most of the system approaches offer not only a theory but also a way of thinking ("systems thinking" or "systemic thinking") and a methodology for dealing with systemic issues or problems.

System dynamics (SD) is a discipline and a methodology for the modeling, simulation and control of complex, dynamic systems. SD was developed by MIT professor Jay W. Forrester (e. g. [20,21]) and has been propagated by his students and associates. SD has grown to a school of numerous academics and practitioners all over the world. The particular approach of SD lies in representing the issues or systems-in-focus as meshes of closed feedback loops made up of stocks and flows, in continuous time and subject to delays.

The development of the system dynamics methodology and the worldwide community that applies SD to modeling and simulation in radically different contexts suggest that it is a "systems approach" on its own. Nevertheless, taking "system dynamics" as the (one and only) synonym for "systemic thinking" would be going too far, given the other approaches to systemic thinking as well as a variety of system theories and methodologies, many of which are complementary to SD. In any case, however, the SD community has become the strongest "school" of the Systems approach, if one takes the numbers of members in organizations representing the different schools as a measure (by 2006, the System Dynamics Society had more than 1000 members).

The rationale and structure of this contribution is as follows. Starting with the emergence of the systems approach, the multiple roots and theoretical streams of systemics are outlined. Next, the common grounds and differences among different strands of the systems approach are highlighted, and the various systems methodologies are explored. Then the distinctive features of SD are analyzed. Finally comes a reflection on the relationships of SD with the rest of the systems movement as well as with potential complementarities and synergies.

In Table 1, a time-line overview of some milestones in the evolution of the systems approach in general and System Dynamics in particular is given. Elaborating on each of the sources quoted therein would reach beyond the purpose of this chapter. However, to convey a synoptic view, a diagram showing the different systems approaches and their interrelationships is provided in the Appendix "Systems Approaches – An Overview".

Emergence of the Systems Approach

The systems movement has many roots and facets, with some of its concepts going back as far as ancient Greece. What we name as "the systems approach" today materialized in the first half of the twentieth century. At least two important components should be mentioned: those proposed by von Bertalanffy and by Wiener.

Ludwig von Bertalanffy, an American biologist of Austrian origin, developed the idea that organized wholes of any kind should be describable and, to a certain extent, explainable, by means of the same categories, and ultimately by one and the same formal apparatus. His *general systems theory* triggered a whole movement which has tried to identify invariant structures and mechanisms across different kinds of organized wholes (for example, hierarchy, teleology, purposefulness, differentiation, morphogenesis, stability, ultrastability, emergence, and evolution).

In 1948 Norbert Wiener, an American mathematician at the Massachusetts Institute of Technology, published his seminal book on *Cybernetics*, building upon interdisciplinary work carried out in cooperation with Bigelow, an IBM engineer, and Rosenblueth, a physiologist. Wiener's opus became the transdisciplinary foundation for a new science of capturing as well as designing control and communication mechanisms in all kinds of dynamic systems [81]. Cyberneticists have been interested in concepts such as information, communication, complexity, autonomy, interdependence, cooperation and conflict, self-production ("autopoiesis"), self-organization, (self-) control, self-reference and (self-) transformation of complex dynamic systems.

Along the genetic line of the tradition which led to the evolution of General Systems Theory (von Bertalanffy, Boulding, Gerard, Miller, Rapoport) and Cybernetics (Wiener, McCulloch, Ashby, Powers, Pask, Beer), a number of roots can be identified, in particular:

- Mathematics (for example, Newton, Poincaré, Lyapunov, Lotka, Volterra, Rashevsky)
- Logic (for example, Epimenides, Leibniz, Boole, Russell and Whitehead, Goedel, Spencer-Brown)
- Biology, including general physiology and neurophysiology (for example, Hippocrates, Cannon, Rosenblueth, McCulloch, Rosen)
- Engineering and computer science, including the respective physical and mathematical foundations (for example, Heron, Kepler, Watt, Euler, Fourier, Maxwell, Hertz, Turing, Shannon and Weaver, von Neumann, Walsh)
- Social and human sciences, including economics (for example, Hume, Adam Smith, Adam Ferguson, John Stuart Mill, Dewey, Bateson, Merton, Simon, Piaget).

In this last-mentioned strand of the systems movement, one focus of inquiry is on the role of feedback in communication and control in (and between) organizations and society, as well as in technical systems. The other focus of interest is on the multidimensional nature and the multilevel structures of complex systems. Specific theory building, methodological developments and pertinent applications have occurred at the following levels:

- Individual and family levels (for example, systemic psychotherapy, family therapy, holistic medicine, cognitive therapy, reality therapy)
- Organizational and societal levels (for example, managerial cybernetics, organizational cybernetics, sociocybernetics, social systems design, social ecology, learning organizations)
- The level of complex (socio-)technical systems (systems engineering)

The notion of "socio-technical systems" has become widely used in the context of the design of organized wholes involving interactions of people and technology (for instance, Linstone's multi-perspectives-framework, known by way of the mnemonic TOP (*T*echnical, *O*rganizational, *P*ersonal/individual).

As can be noted from these preliminaries, different kinds of system theory and methodology have evolved over time. One of these is a theory of dynamic systems by Jay W. Forrester, which serves as a basis for the methodology of system dynamics. Two eminent titles are [20] and [21]. In SD, the main emphasis falls on the role of structure and its relationship with the dynamic behavior of systems, modeled as networks of informationally closed feedback loops between stock and flow variables. Several other mathematical systems theories have been elaborated, for example, mathematical general systems theory (Klir, Pestel, Mesarovic and Takahara), as well as a whole stream of theoretical developments which can be subsumed under the terms "dynamic systems theory" or "theories of nonlinear dynamics" (for example, catastrophe theory, chaos theory and complexity theory). Under the latter, branches such as the theory of fractals (Mandelbrot), geometry of behavior (Abraham), self-organized criticality (Bak), and network theory (Barabasi, Watts) are subsumed. In this context, the term "sciences of complexity" is used.

In addition, a number of mathematical theories, which can be called "system theories," have emerged in different application contexts, examples of which are discernible in the following fields:

- Engineering, namely information and communication theory (Shannon and Weaver), technology and computer-aided systems theory (for example, control theory, automata, cellular automata, agent-based modeling, artificial intelligence, cybernetic machines, neural nets)
- Operations research (for example, modeling theory and simulation methodologies, Markov chains, genetic algorithms, fuzzy control, orthogonal sets, rough sets)
- Social sciences, economics in particular (for example, game theory, decision theory)
- Biology (for example, Sabelli's Bios theory of creation)
- Ecology (for example, E. and H. Odum's systems ecology).

Most of these theories are transdisciplinary in nature, i. e., they can be applied across disciplines. The Bios theory, for example is applicable to clinical, social, ecological and personal settings [54]. Examples of essentially non-mathematical system theories can be found in many different areas of study, e. g.:

- Economics, namely its institutional/evolutionist strand (Veblen, Myrdal, Boulding, Dopfer)
- Sociology (for example, Parsons' and Luhmann's social system theories, Hall's cultural systems theory)
- Political sciences (for example, Easton, Deutsch, Wallerstein)
- Anthropology (for example, Levi Strauss's structuralist-functionalist anthropology, Margaret Mead)
- Semiotics (for example, general semantics (Korzybski, Hayakawa, Rapoport), cybersemiotics (Brier))
- Psychology and psychotherapy (for example, systemic intervention (Bateson, Watzlawick, F. Simon), and fractal affect logic (Ciompi))

- Ethics and epistemology (for example, Vickers, Churchman, von Foerster, van Gigch)

Several system-theoretic contributions have merged the quantitative and the qualitative in new ways. This is the case for example in Rapoport's works in game theory as well as general systems theory, Pask's conversation theory, von Foerster's cybernetics of cybernetics (second-order cybernetics), and Stafford Beer's opus in managerial cybernetics. In all four cases, mathematical expression is virtuously connected to ethical, philosophical, and epistemological reflection. Further examples are Prigogine's theory of dissipative structures, Mandelbrot's theory of fractals, complex adaptive systems (Holland et al.), Kauffman's complexity theory, and Haken's synergetics, all of which combine mathematical analysis and a strong component of qualitative interpretation.

A large number of systems methodologies, with the pertinent threads of systems practice, have emanated from these theoretical developments. Many of them are expounded in detail in specialized encyclopedias (e. g., [27] and, under a specific theme, named *Systems Science and Cybernetics*, of the Encyclopedia of Life Support Systems [18]). In this chapter, only some of these will be addressed explicitly, in order to shed light on the role of SD as part of the systems movement.

Common Grounds and Differences

Even though the spectrum of system theories and methodologies outlined in the preceding section may seem multifarious, all of them have a strong common denominator: They build on the idea of systems as organized wholes. An objectivist working definition of a system is that of a whole, the organization of which is made up by interrelationships. A subjectivist definition is that of a set of interdependent variables in the mind of an observer, or, a mental construct of a whole, an aspect that has been emphasized by the position of constructivism. *Constructivism* is a synonym for *second-order cybernetics*. While first-order cybernetics concentrates on regulation, information and feedback, second-order cybernetics focuses on observation, self-organization and self-reference. Heinz von Foerster established the distinction between 'observed systems' for the former and 'observing systems' for the latter [74].

From the standpoint of operational philosophy, a system is, as Rapoport says, "a part of the world, which is sufficiently well defined to be the object of an inquiry or also something, which is characterized by a structure, for example, a production system" [50].

In recent systems theory, the aspect of relationships has been emphasized as the main building block of a system, as one can see from a definition published by the International Society for the Systems Sciences (ISSS): "A system is a family of relationships between its members acting as a whole" [63]. Also, purpose and interaction have played an important part in reflections on systems: Systems are conceived, in the words of Forrester [21], as "wholes of elements, which cooperate towards a common goal." Purposeful behavior is driven by internal goals, while purposive behavior rests on a function assigned from the outside. Finally, the aspects of open and closed functioning have been emphasized. Open systems are characterized by the import and export of matter, energy and information. A variant of particular relevance in the case of social systems is the operationally closed system, that is, a system which is self-referential in the sense that its self-production (autopoiesis) is a function of production rules and processes by which order and identity are maintained, and which cannot be modified directly from outside. As we shall see, this concept of operational closure is very much in line with the concept of circularity used in SD.

At this point, it is worth elaborating on the specific differences between two major threads of the systems movement, which are of special interest because they are grounded in "feedback thought" [52]: The cybernetic thread, from which organizational cybernetics has emanated, and the servomechanic thread in which SD is grounded. As Richardson's detailed study shows, the strongest influence on cybernetics came from biologists and physiologists, while the thinking of economists and engineers essentially shaped the servomechanic thread. Consequently, the concepts of the former are more focused on the adaptation and control of complex systems for the purpose of maintaining stability under exogenous disturbances. Servomechanics, on the other hand, and SD in particular, take an endogenous view, being mainly interested in understanding circular causality as the principal source of a system's behavior. Cybernetics is more connected with communication theory, the general concern of which can be summarized as how to deal with randomly varying input. SD, on the other hand, shows a stronger link with engineering control theory, which is primarily concerned with behavior generated by the control system itself, and by the role of nonlinearities. Managerial cybernetics and SD both share the concern of contributing to management science, but with different emphases and with instruments that are different but in principle complementary. Finally, the mathematical foundations are generally more evident in the basic literature on SD than in the writings on organizational cybernetics, in which the formal apparatus underlying model formulation is confined to a small number of publications (e. g., [7,10]), which are less known than the qualitative treatises. The terms *management cybernetics* and *managerial cybernetics* are used as synonyms for *organizational cybernetics*.

The Variety of Systems Methodologies

The methodologies that have evolved as part of the systems movement cannot be expounded in detail here. The two epistemological strands in which they are grounded, however, can be identified – the positivist tradition and the interpretivist tradition.

Positivist tradition denotes those methodological approaches that focus on the generation of "positive knowledge," that is, a knowledge based on "positively" ascertained facts. *Interpretivist tradition* denotes those methodological approaches that emphasize the importance of subjective interpretations of phenomena. This stream goes back to Greek art and science of the interpretation and understanding of texts.

Some systems methodologies have been rooted in the positivist tradition, and others in the interpretivist tradition. The differences between the two can be described along the following set of polarities:

- An objectivist versus a subjectivist position
- A conceptual–instrumental versus a communicational/cultural/political rationality
- An inclination to quantitative versus qualitative modeling
- A structuralist versus a discursive orientation.

A positivistic methodological position tends toward the objectivistic, conceptual–instrumental, quantitative and structuralist–functionalist in its approach. An interpretive position, on the other hand, tends to emphasize the subjectivist, communicational, cultural, political, ethical and esthetic—that is, the qualitative and discursive aspects. It would be too simplistic to classify a specific methodology in itself as being "positivistic" or "interpretative". Despite the traditions they have grown out of, several methodologies have evolved and been reinterpreted or opened to new aspects (see below).

In the following, a sample of systems methodologies will be characterized and positioned in relation to these two traditions, beginning with those in the positivistic strand:

- *"Hard" OR methods*. Operations research (OR) uses a wide variety of mathematical and statistical methods and techniques—for example of optimization, queuing, dynamic programming, graph theory, time series analysis—to provide solutions for organizational and managerial problems, mainly in the operational domains of production and logistics, and in finance.
- *Living systems theory*. In his LST, James Grier Miller [44] identifies a set of 20 necessary components that can be discerned in living systems of any kind. These structural features are specified on the basis of a huge empirical study and proposed as the "critical subsystems" that "make up a living system." LST has been used as a device for diagnosis and design in the domains of engineering and the social sciences.
- *Viable system model*. To date, Stafford Beer's VSM is probably the most important product of organizational cybernetics. It specifies a set of management functions and their interrelationships as the sufficient conditions for the viability of any human or social system (see [10]). These are applicable in a recursive mode, for example, to the different levels of an organization. The VSM has been widely applied in the diagnostic mode, but also to support the design of all kinds of social systems. Specific methodologies for these purposes have been developed, for instance for use in consultancy. The term viable system diagnosis (VSD) is also used.

The methodologies and models addressed up to this point have by and large been created in the positivistic tradition of science. Other strands in this tradition do exist, e. g., systems analysis and systems engineering, which together with OR have been called "hard systems thinking" (p. 127 in [31]). Also, more recent developments such as mathematical complexity and network theories, agent-based modeling and most versions of game theory can be classified as hard systems approaches.

The respective approaches have not altogether been excluded from fertile contacts with the interpretivist strand of inquiry. In principle, all of them can be considered as instruments for supporting discourses about different interpretations of an organizational reality or alternative futures studied in concrete cases. In our time, most applications of the VSM, for example, are constructivist in nature. To put it in a nutshell, these applications are (usually collective) constructions of a (new) reality, in which observation and interpretation play a crucial part. In this process, the actors involved make sense of the system under study, i. e., the organization in focus, by mapping it on the VSM. At the same time they bring forth "multiple realities rather than striving for a fit with one reality" (p. 299 in [29]).

The second group of methodologies is part of the interpretive strand:

- *Interactive Planning*. IP is a methodology, designed by Russell Ackoff [1], and developed further by Jamshid

Gharajedaghi [28], for the purpose of dealing with "messes" and enabling actors to design their desired futures, as well as to bring them about. It is grounded in theoretical work on purposeful systems, reverts to the principles of continuous, participative and holistic planning, and centers on the idea of an "idealized design."

- *Soft Systems Methodology*. SSM is a heuristic designed by Peter Checkland [13,14] for dealing with complex situations. Checkland suggests a process of inquiry constituted by two aspects: A conceptual one, which is logic based, and a sociopolitical one, which is concerned with the cultural feasibility, desirability and implementation of change.
- *Critical Systems Heuristics*. CSH is a methodology, which Werner Ulrich [67,68] proposed for the purpose of scientifically informing planning and design in order to lead to an improvement in the human condition. The process aims at uncovering the interests that the system under study serves. The legitimacy and expertise of actors, and particularly the impacts of decisions and behaviors of the system on others – the "affected" – are elicited by means of a set of boundary questions. CSH can be seen as part of a wider movement known as the "Emancipatory Systems Approach" which embraces, e. g., Freire's Critical Pedagogy, Interpretive Systemology, and Community OR (see pp. 291ff in [31]).

All three of these methodologies (IP, SSM, and CSH) are positioned in the interpretive tradition. Other methodologies and concepts which can be subsumed under the interpretive systems approach are, e. g., Warfield's science of generic design, Churchman's social system design, Senge's soft systems thinking, Mason and Mitroff's strategic assumptions surfacing and testing (SAST), Eden and Ackermann's strategic options in development and analysis (SODA), and other methodologies of soft operational research (for details, see pp. 211ff in [31]). The interpretive methodologies were designed to deal with qualitative aspects in the analysis and design of complex systems, emphasizing the communicational, social, political and ethical dimensions of problem solving. Several authors mention explicitly that they do not preclude the use of quantitative techniques or include such techniques in their repertoire (e. g., the biocyberneticist Frederic Vester).

In an advanced understanding of system dynamics both of these traditions—positivist and interpretivist—are synthesized. The adherents of SD conceive of model building and validation as a semi-formal, relativistic, holistic social process. Validity is understood as usefulness or fitness in relation to the purpose of the model, and validation as an elaborate set of procedures – including logico-structural, heuristic, algorithmic, statistical, and also discursive components – by which the quality of and the confidence in a model are gradually improved (see [4,5,59]).

System Dynamics – Its Features, Strengths and Limitations

The features, strengths and limitations of the SD methodology are a consequence of its specific characteristics. In the context of the multiple theories and methodologies of the systems movement, some of the distinctive features of SD are (for an overview, see [52], pp. 142ff in [31]):

- *Feedback as conceptual basis*. SD model systems are high-order, multiple-loop networks of closed loops of information. Concomitantly, an interest in non-linearities, long-term patterns and internal structure rather than external disturbances is characteristic of SD (p. 31 in [40]). However, SD models are not "closed systems", as sometimes is claimed, in the sense that (a) flows can originate from outside the system's boundaries, (b) representations of exogenous factors or systems can be incorporated into any model as parameters or special modules, and (c) new information can be accommodated via changes to a model. In other words, the SD view hinges on a view of systems which are closed in a causal sense but not materially (p. 297 in [52]).
- *Focus on internally generated dynamics*. SD models are conceived as closed systems. The interest of users is in the dynamics generated inside those systems. Given the nature of closed feedback loops and the fact that delays occur within them, the dynamic behavior of these systems is essentially non-linear.
- *Emphasis on understanding*. For system dynamicists the understanding of the dynamics of a system is the first goal to be achieved by means of modeling and simulation. Conceptually, they try to understand events as embedded in patterns of behavior, which in turn are generated by underlying structures. Such understanding is enabled by SD as it "shows how present policies lead to future consequences" (Sect. VIII in [23]). Thereby, the feedback loops are "a major source of puzzling behavior and policy difficulties" (p. 300 in [52]). SD models purport to test mental models, hone intuition and improve learning (see [65]).
- *High degree of operationality*. SD relies on formal modeling. This fosters disciplined thinking; assumptions, underlying equations and quantifications must be clarified. Feedback loops and delays are visualized and formalized; therewith the causal logic inherent in a model

is made more transparent and discussable than in most other methodologies [53]. Also, a high level of realism in the models can be achieved. SD is therefore apt to support decision-making processes effectively.

- *Far-reaching requirements (and possibilities) for the combination of qualitative and quantitative aspects of modeling and simulation*. This is a consequence of the emphasis on understanding. The focus is not on point-precise prediction, but on the generation of insights into the patterns generated by the systems under study.
- *High level of generality and scale robustness*. The representation of dynamic systems in terms of stocks and flows is a generic form, which is adequate for a wide spectrum of potential applications. This spectrum is both broad as to the potential subjects under study, and deep as to the possible degrees of resolution and detail [38]. In addition, the SD methodology enables one to deal with large numbers of variables within multiple interacting feedback loops (p. 9 in [22]). SD has been applied to the most diverse subject areas, e. g., global modeling, environmental issues, social and economic policy, corporate and public management, regional planning, medicine, psychology and education in mathematics, physics and biology.

The features of SD just sketched out result in both strengths and limitations. We start with the strengths.

Strengths of SD

1. Its *specific modeling approach* makes SD particularly helpful in gaining insights into the patterns exhibited by dynamic systems, as well as the structures underlying them. Closed-loop modeling has been found most useful in fostering understanding of the dynamic functioning of complex systems. Such understanding is especially facilitated by the principle of modeling the systems or issues under study in a continuous mode and at rather high aggregation levels [20,38]. With the help of relatively small but insightful models, and by means of sensitivity analyses as well as optimization heuristics incorporated in the application software packages, decision-spaces can be thoroughly explored. Vulnerabilities and the consequences of different system designs can be examined with relative ease.
2. The *generality of the methodology* and its power to crystallize operational thinking in realistic models have triggered applications in the most varied contexts. Easy-to-use software and the features of screen-driven modeling via graphic user interfaces provide a strong lever for collaborative model-building in teams (cf. [2,69]).
3. Another strong point is the *momentum of the SD movement*. Due to the strengths commented above this point, the community of users has grown steadily, being probably the largest community within the systems movement. Lane (p. 484 in [36]) has termed SD "one of the most widely used systems approaches in the world."
4. Its specific features make SD an exceptionally effective tool for *conveying systemic thinking* to anybody. Therefore, it also has an outstanding track-record of classroom applications for which "learner-directed learning" [24] or "learner-centered learning" is advocated [25,26]. Pertinent audiences range from schoolchildren at the levels of secondary and primary schools to managers and scientists.

Given these strengths, the community of users has not only grown significantly, but has also transcended disciplinary boundaries, ranging from the formal and natural sciences to the humanities, and covering multiple uses from theory building and education to the tackling of real-world problems at almost any conceivable level. Applications to organizational, societal and ecological issues have seen a particularly strong growth. This feeds back on the availability and growth of the knowledge upon which the individual modeler can draw.

The flip side of most of the strengths outlined here embodies the limitations of SD; we concentrate on those which can be relevant to a possible complementarity of SD with other systems methodologies.

Limitations of SD

1. The main point here is that SD does not provide a framework or methodology for the *diagnosis and design* of organizational structures in the sense of interrelationships among organizational actors. This makes SD susceptible to completion from without – a completion which organizational cybernetics (OC), and the VSM in particular, but also living system theory (LST), especially can provide. The choice falls on these two approaches because of their strong heuristic power and their complementary strengths in relation to SD (cf. [57,61]).
2. Another limitation of SD is related to the *absorption of variety* (complexity) by an organization. *Variety* is a technical term for *complexity*, which denotes a (high) number of potential states or behaviors of a system (based on [3,8]). SD offers an approach to the handling of variety which allows modeling at different scales of a problem or system [47]. It focuses on the identification, at a certain resolution level or possibly several resolution levels, of the main stock variables which will be

affected by the respective flows. These, in turn, will be influenced by parameters and auxiliary variables. This approach, even though it enables thinking and modeling at different scales, does not provide a formal procedure for an organization to cope with the external complexity it faces, namely, for designing a structure which can absorb that complexity. In contrast, OC and LST offer elaborate models to enable the absorption of variety, in the case of the VSM based explicitly on Ashby's *Law of Requisite Variety*. It says "Only variety can destroy variety", which implies that the varieties of two interacting systems must be in balance, if stability is to be achieved [3]. The VSM has two salient features in this respect. Firstly, it helps design an organizational unit for viability, by enabling it to attenuate the complexity of its environment, and also to enhance its eigenvariety, so that the two are in balance. The term *variety engineering* has been used in this context [9]. Secondly, the recursive structure of the VSM ensures that an organization with several levels will develop sufficient eigenvariety along the fronts on which the complexity it faces unfolds. Similarly, LST offers the conditions for social systems to survive, by maintaining thermodynamically highly improbable energy states via continuous interaction with their environments. The difference between the two approaches is that the VSM functions more in the strategic and informational domains, while the LST model essentially focuses on the operational domain. In sum, both can make a strong contribution related to coping with the external complexity faced by organizations, and therefore can deliver a strong complement to SD.

3. Finally, the design of *modeling processes* confronts SD with specific challenges. The original SD methodology of modeling and simulation was to a large extent functionally and technically oriented. This made it strong in the domain of logical analysis, while the socio-cultural and political dimensions of the modeling process were, if not completely out of consideration, at least not a significant concern in methodological developments. The SD community – also under the influence of the soft systems approaches – has become aware of this limitation and has worked on incorporating features of the social sciences into its repertoire. The following examples, which document this effort to close the gap, stand for many. Extensive work on group model building has been achieved, which explores the potential of collaborative model building [69]. A new schema for the modeling process has been proposed, which complements logic-based analysis by cultural analysis [37]. The social dimension of system dynamics-based modeling has become subject to intensive discussion ([77]; and other contributions to the special issue of *Systems Research and Behavioral Science*, Vol. 51, No. 4, 2006). Finally, in relation to consultancy methodology, modeling has been framed as a learning process [34] and as second-order intervention [60].

As has been shown, there is a need to complement classical SD with other methodologies, when issues are at stake which it cannot handle by itself. VSM and LST are excellent choices when issues of organizational diagnosis or design are to be tackled.

The limitations addressed here call attention to other methodologies which exhibit certain features that traditionally were not incorporated, or at least not explicit, in SD methodology. One aspect concerns the features that explicitly address the subjectivity of purposes and meanings ascribed to systems. In this context, support for problem formulation, model construction and strategy design by individuals on the one hand and groups on the other are relevant issues. Also, techniques for an enhancement of creativity (e. g., the generation and the reframing of options) in both individuals and groups are a matter of concern. Two further aspects relate to methodological arrangements for coping with the specific issues of negotiation and alignment in pluralist and coercive settings.

As far as the modeling processes are concerned, group model building has proven to be a valuable complement to pure modeling and simulation. However, there are other systems methodologies which should be considered as potentially apt to enrich SD analysis, namely the soft approaches commented upon earlier, e. g., interactive planning, soft system methodology and critical system heuristics.

On the other hand, SD can be a powerful complement to other methodologies which are more abstract or more static in nature. This potential refers essentially to all systems approaches which stand in the interpretive ("soft") tradition, but also to approaches which stand in the positivist traditions, such as the VSM and LST. These should revert to the support of SD in the event that tradeoffs between different goals must be handled, or if implications of long-term decisions on short-term outcomes (and vice versa) have to be ascertained, and whenever contingencies or vulnerabilities must be assessed.

Actual and Potential Relationships

It should be clear by now that the systems movement has bred a number of theories and methodologies, none of which can be considered all-embracing or complete. All of them have their strengths and weaknesses, and their specific potentials and limitations.

Since Burrell and Morgan [12] adverted to incommensurability between different paradigms of social theory, several authors have acknowledged or even advocated methodological complementarism. They argue that there is a potential complementarity between different methods, and, one may add, models, even if they come from distinct paradigms. Among these authors are, e. g., Brocklesby [11], Jackson [30], Midgley [43], Mingers [45], Schwaninger [55] and Yolles [83]. These authors have opened up a new perspective in comparison with the non-complementaristic state-of-the-art.

In the past, the different methodologies have led to the formation of their own traditions and "schools," with boundaries across which not much dialogue has evolved. The methodologies have kept their protagonists busy testing them and developing them further. Also, the differences between different language games and epistemological traditions have often suggested incommensurability, and therewith have impaired communication. Prejudices and a lack of knowledge of the respective other side have accentuated this problem: Typically, "hard" systems scientists are suspicious of "soft" systems scientists. For example, many members of the OR community, not unlike orthodox quantitatively oriented economists, adhere to the opinion that "SD is too soft." On the other hand the protagonists of "soft" systems approaches, even though many of them have adopted feedback diagrams (causal loop diagrams) for the sake of visualization, are all too often convinced that "SD is too hard." Both of these judgments indicate a lack of knowledge, in particular of the SD validation and testing methods available, on the one hand, and the technical advancements achieved in modeling and simulation, on the other (see [5,59,66]).

In principle, both approaches are complementary. The qualitative view can enrich quantitative models, and it is connected to their philosophical, ethical and esthetical foundations. However, qualitative reasoning tends to be misleading if applied to causal network structures without being complemented by formalization and quantification of relationships and variables. Furthermore, the quantitative simulation fosters insights into qualitative patterns and principles. It is thus a most valuable device for validating and honing the intuition of decision makers, via corroboration and falsification.

Proposals that advocate mutual learning between the different "schools" have been formulated inside the SD community (e. g., [35]). The International System Dynamics Conference of 1994 in Stirling, held under the banner of "Transcending the Boundaries," was dedicated to the dialogue between different streams of the systems movement.

Also, from the 1990s onwards, there were vigorous efforts to deal with methodological challenges, which traditionally had not been an important matter of scientific interest within the SD community. Some of the progress made in these areas is documented in a special edition of *Systems Research and Behavioral Science* (Vol. 21, No. 4, July-August 2004). The main point is that much of the available potential is based on the complementarity, not the mutual exclusiveness, of the different systems approaches.

In the future, much can be gained from leveraging these complementarities. Here are two examples of methodological developments in this direction, which appear to be achievable and potentially fertile: The enhancement of qualitative components in "soft" systems methodologies in the process of knowledge elicitation and model building (cf. [69]), and the combination of cybernetics-based organizational design with SD-based modeling and simulation (cf. [61]). Potential complementarities exist not only across the qualities – quantities boundary, but also within each one of the domains. For example, with the help of advanced software, SD modeling ("top-down") and agent-based modeling ("bottom-up") can be used in combination.

From a meta-methodological stance, generalist frameworks have been elaborated which contain blueprints for combining different methodologies where this is indicated. Two examples are:

- *Total systems intervention* (TSI) is a framework proposed by Flood and Jackson [19], which furnishes a number of heuristic schemes and principles for the purpose of selecting and combining systems methods/methodologies in a customized way, according to the issue to be tackled. SD is among the recommended "tools".
- *Integrative systems methodology* (ISM) is a heuristic for providing actors in organizations with requisite variety, developed by Schwaninger [55,56]. It advocates (a) dealing with both content– and context-related issues during the process, and (b) placing a stronger emphasis on the validation of qualitative and quantitative models as well as strategies, in both dimensions of the content of the issue under study and the organizational context into which that issue is embedded. For this purpose, the tools of SD (to model content) and organizational cybernetics – the VSM (to model context) – are cogently integrated.

These are only two examples. In principle, SD could make an important contribution in the context of most of the methodological frameworks, far beyond the extent to which this has been the case. Systems methodologists and practitioners can potentially benefit enormously from including SD methodology in their repertoires.

Outlook

There have recently been calls for an eclectic "mixing and matching" of methodologies. In light of the epistemological tendencies of our time towards radical relativism, it is necessary to warn against taking a course in which "anything goes". It is most important to emphasize that the desirable methodological progress can only be achieved on the grounds of scientific rigor. This postulate of "rigor" is not to be confused with an encouragement of "rigidity." The necessary methodological principles advocated here are disciplined thinking, a permanent quest for better models (that is, thorough validation), and the highest achievable levels of transparency in the formalizations as well as of the underlying assumptions and sources used. Scientific rigor, in this context, also implies that combinations of methodologies reach beyond merely eclectic addons from different methodologies, so that genuine integration towards better adequacy to the issues at hand is achieved.

The contribution of system dynamics can come in the realms of the following:

- Fostering disciplined thinking
- Understanding dynamic behaviors of systems and the structures that generate them
- Exploring paths into the future and the concrete implications of decisions
- Assessing strategies as to their robustness and vulnerabilities, in ways precluded by other, more philosophical, and generally "soft" systems approaches

These latter streams can contribute to reflecting and tackling the meaning- and value-laden dimensions of complex human, social and ecological systems. Some of their features should and can be combined synergistically with system dynamics, particularly by being incorporated into the repertoires of system dynamicists. From the reverse perspective, incorporating system dynamics as a standard tool will be of great benefit for the broad methodological frameworks. Model formalization and dynamic simulation may even be considered necessary components for the study of the concrete dynamics of complex systems.

Finally, there are also many developments in the "hard", i. e., mathematics-, statistics-, logic-, and informatics-based methods and technologies, which are apt to enrich the system dynamics methodology, namely in terms of modeling and decision support. For example, the constantly evolving techniques of time-series analysis, filtering, neural networks and control theory can improve the design of system-dynamics-based systems of (self-)control. Also, a bridge across the divide between the top-down modeling approach of SD and the bottom-up approach of agent-based modeling appears to be feasible. Furthermore, a promising perspective for the design of genuinely "intelligent organizations" emerges if one combines SD with advanced database-management, cooperative model building software, and the qualitative features of the "soft" systems methodologies.

The approaches of integrating complementary methodologies outlined in this contribution definitely mark a new phase in the history of the systems movement.

Appendix

Milestones in the Evolution of the Systems Approach in General and System Dynamics in Particular

The table gives an overview of the systems movement's evolution, as shown in its main literature; and that overview is not exhaustive.

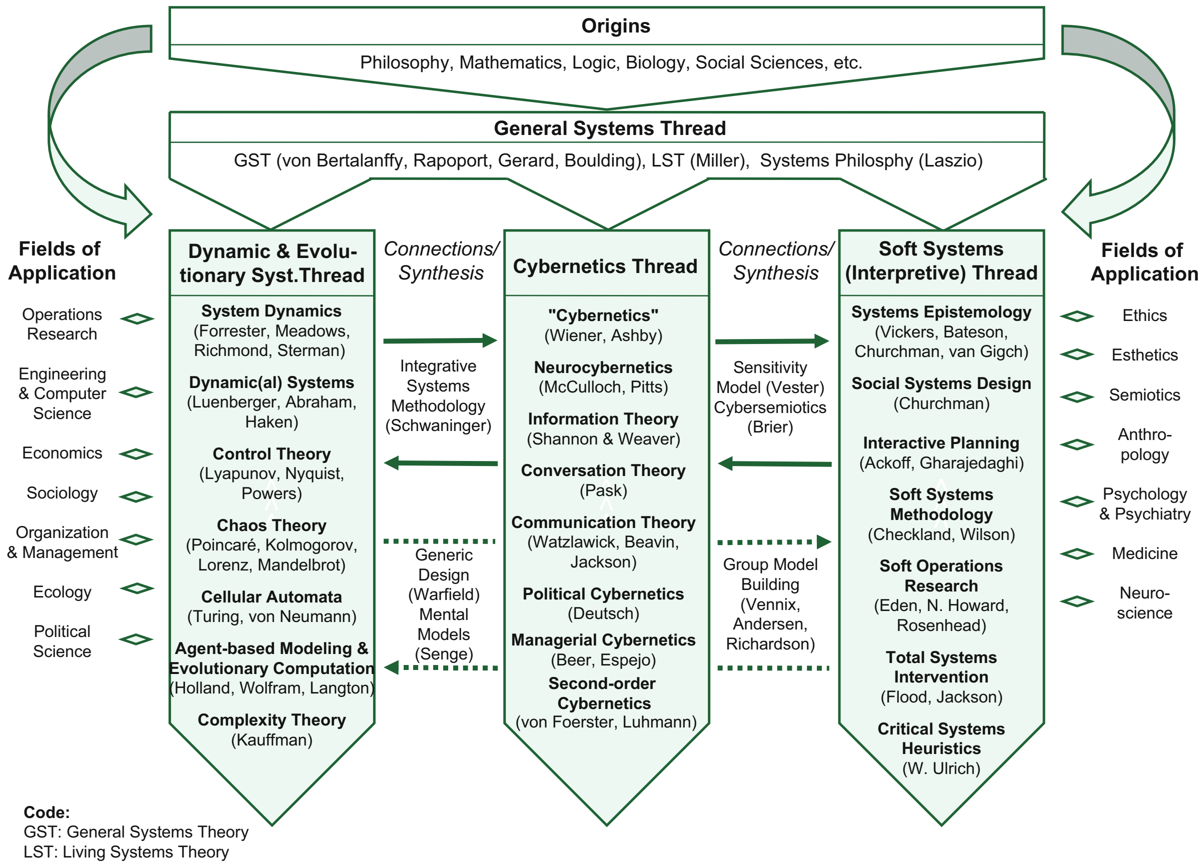
Systems Approaches – An Overview

Note: This diagram shows three streams of the systems approach in the context of their antecedents. The general systems thread has its origins in philosophical roots from antiquity: The term *system* derives from the old Greek *σύστημα* (systēma), while, *cybernetics* stems from the Greek *κυβερνήτης* (kybernētēs). The arrows between the threads stand for interrelationships and efforts to synthesize the connected approaches. For example, integrated systems methodology is an integrative attempt to leverage the complementarities of system dynamics and organizational cybernetics. Enumerated to the left and right of the scheme are the fields of application. The big arrows in the upper region of the diagram indicate that the roots of the systems approach continue influencing the different threads and the fields of application even if the path via general systems theory is not pursued.

The diagram is not a complete representation, but the result of an attempt to map the major threads of the systems movement and some of their interrelations. Hence, the schema does not cover all schools or protagonists of the movement. Why does the diagram show a dynamic and evolutionary systems thread and a cybernetics thread, if cybernetics is about dynamic systems? The latter embraces all the approaches that are explicitly grounded in cybernetics. The former relates to all other approaches

System Dynamics in the Evolution of the Systems Approach, Table 1
Milestones in the evolution of the systems approach in general and system dynamics in particular

Foundations of general system theory		
Von Bertalanffy	Zu einer allgemeinen Systemlehre	1945
	An Outline of General System Theory	1950
	General System Theory	1968
Bertalanffy, Boulding, Gerard, Rapoport	Foundation of the Society for General Systems Research	1953
Klir	An Approach to General System Theory	1968
Simon	The Sciences of the Artificial	1969
Pichler	Mathematische Systemtheorie	1975
Miller	Living Systems	1978
Mesarovic & Takahara	Abstract Systems Theory	1985
Rapoport	General System Theory	1986
Foundations of cybernetics		
Macy Conferences (Josiah Macy, Jr. Foundation)	Cybernetics. Circular Causal, and Feedback Mechanisms in Biological and Social Systems	1946–1951
Wiener	Cybernetics or Control and Communication in the Animal and in the Machine	1948
Ashby	An Introduction to Cybernetics	1956
Pask	An Approach to Cybernetics	1961
Von Foerster, Zopf	Principles of Self-Organization	1962
McCulloch	Embodiments of Mind	1965
Foundations of organizational cybernetics		
Beer	Cybernetics and Management	1959
	Towards the Cybernetic Factory	1962
	Decision and Control	1966
	Brain of the Firm	1972
Von Foerster	Cybernetics of Cybernetics	1974
Foundations of system dynamics		
Forrester	Industrial Dynamics	1961
	Principles of Systems	1968
	Urban Dynamics	1969
	World Dynamics	1971
Meadows et al.	Limits to Growth	1972
Richardson	Feedback Thought in Social Science and Systems Theory	1991
Systems methodology		
Churchman	Challenge to Reason	1968
	The Systems Approach	1968
Vester & von Hesler	Sensitivitätsmodell	1980
Checkland	Systems Thinking, Systems Practice	1981
Ackoff	Creating the Corporate Future	1981
Ulrich	Critical Heuristics of Social Planning	1983
Warfield	A Science of Generic Design	1994
Schwaninger	Integrative Systems Methodology	1997
Gharajedaghi	Systems Thinking	1999
Sabelli	Bios – A Study of Creation	2005
Selected recent works in system dynamics		
Senge	The Fifth Discipline	1990
Barlas & Carpenter	Model Validity	1990
Vennix	Group Model Building	1996
Lane & Oliva	Synthesis of System Dynamics and Soft Systems Methodology	1998
Sterman	Business Dynamics	2000
Warren	Strategy Dynamics	2002, 2008
Wolstenholme	Archetypal Structures	2003
Morecroft	Strategic Modelling	2007
Schwaninger & Grösser	Theory-building with System Dynamics & Model Validation	2008, 2009


Origins
Philosophy, Mathematics, Logic, Biology, Social Sciences, etc.
General Systems Thread
GST (von Bertalanffy, Rapoport, Gerard, Boulding), LST (Miller), Systems Philosphy (Laszio)
Fields of Application
Operations Research
Engineering & Computer Science
Economics
Sociology
Organization & Management
Ecology
Political Science
Dynamic & Evolutionary Syst.Thread
System Dynamics (Forrester, Meadows, Richmond, Sterman)
Dynamic(al) Systems (Luenberger, Abraham, Haken)
Control Theory (Lyapunov, Nyquist, Powers)
Chaos Theory (Poincaré, Kolmogorov, Lorenz, Mandelbrot)
Cellular Automata (Turing, von Neumann)
Agent-based Modeling & Evolutionary Computation (Holland, Wolfram, Langton)
Complexity Theory (Kauffman)
Connections/ Synthesis
Integrative Systems Methodology (Schwaninger)
Generic Design (Warfield) Mental Models (Senge)
Cybernetics Thread
"Cybernetics" (Wiener, Ashby)
Neurocybernetics (McCulloch, Pitts)
Information Theory (Shannon & Weaver)
Conversation Theory (Pask)
Communication Theory (Watzlawick, Beavin, Jackson)
Political Cybernetics (Deutsch)
Managerial Cybernetics (Beer, Espejo)
Second-order Cybernetics (von Foerster, Luhmann)
Connections/ Synthesis
Sensitivity Model (Vester) Cybersemiotics (Brier)
Group Model Building (Vennix, Andersen, Richardson)
Soft Systems (Interpretive) Thread
Systems Epistemology (Vickers, Bateson, Churchman, van Gigch)
Social Systems Design (Churchman)
Interactive Planning (Ackoff, Gharajedaghi)
Soft Systems Methodology (Checkland, Wilson)
Soft Operations Research (Eden, N. Howard, Rosenhead)
Total Systems Intervention (Flood, Jackson)
Critical Systems Heuristics (W. Ulrich)
Fields of Application
Ethics
Esthetics
Semiotics
Anthropology
Psychology & Psychiatry
Medicine
Neuroscience
Code:
GST: General Systems Theory
LST: Living Systems Theory

System Dynamics in the Evolution of the Systems Approach, Figure 1

concerned with dynamic or evolutionary systems. The simplification made it necessary to somewhat curtail logical perfection for the sake of conveying a synoptic view of the different systems approaches, in a language that uses the categories common in current scientific and professional discourse. Overlaps exist, e. g., between dynamic systems and chaos theory, cellular automata and agent-based modeling.

Bibliography

Primary Literature

1. Ackoff RL (1981) Creating the Corporate Future. Wiley, New York
2. Andersen DF, Richardson GP (1997) Scripts for Group Model Building. Syst Dyn Rev 13(2):107–129
3. Ashby WR (1956) An Introduction to Cybernetics. Chapman & Hall, London
4. Barlas Y (1996) Formal aspects of model validity and validation in system dynamics. Syst Dyn Rev 12(3):183–210
5. Barlas Y, Carpenter S (1990) Philosophical roots of model validity: Two paradigms. Syst Dyn Rev 6(2):148–166
6. Beer S (1959) Cybernetics and Management. English Universities Press, London
7. Beer S (1994) Towards the Cybernetic Factory. In: Harnden R, Leonard A (eds) How Many Grapes Went into the Wine. Stafford Beer on the Art and Science of Holistic Management. Wiley, Chichester, pp 163–225 (reprint, originally published in 1962)
8. Beer S (1966) Decision and Control. Wiley, Chichester
9. Beer S (1979) The Heart of Enterprise. Wiley, Chichester
10. Beer S (1981) Brain of the Firm, 2nd edn. Wiley, Chichester
11. Brocklesby J (1993) Methodological complementarism or separate paradigm development – Examining the options for enhanced operational research. Aust J Manag 18(2):133–157
12. Burrell G, Morgan G (1979) Sociological Paradigms and Organisational Analysis. Hants, Gower
13. Checkland PB (1981) Systems Thinking, Systems Practice. Wiley, Chichester
14. Checkland PB, Poulter J (2006) Learning for Action: A Short Definitive Account of Soft Systems Methodology, and its Use Practitioners, Teachers and Students. Wiley, Chichester
15. Churchman CW (1968) Challenge to Reason. McGraw-Hill, New York
16. Churchman CW (1968) The Systems Approach. Delacorte Press, New York
17. Churchman CW (1979) The Systems Approach and its Enemies. Basic Books, New York
18. Encyclopedia of Life Support Systems (2002) published under: http://www.eolss.net/

19. Flood RL, Jackson MC (1991) Creative Problem Solving. Total Systems Intervention. Wiley, Chichester
20. Forrester JW (1961) Industrial Dynamics. MIT Press, Cambridge
21. Forrester JW (1968) Principles of Systems. MIT Press, Cambridge
22. Forrester JW (1969) Urban Dynamics. MIT Press, Cambridge
23. Forrester JW (1971) World Dynamics. Pegasus Communications, Waltham
24. Forrester JW (1993) System Dynamics and the Lessons of 35 Years. In: DeGreene KB (ed) Systems-Based Approach to Policy Making. Kluwer, Boston
25. Forrester JW (1993) System Dynamics as an Organizing Framework for Pre-college Education. Syst Dyn Rev 9(2):183–194
26. Forrester JW (1997) System Dynamics and K-12 Teachers. A Lecture at the University of Virginia School of Education, Massachusetts Institute of Technology. System Dynamics Group Paper D-4665–4
27. François C (2004) International Encyclopedia of Systems and Cybernetics, 2nd edn. Saur, München
28. Gharajedaghi J (1999) Systems Thinking. Managing Chaos and Complexity. Butterworth-Heinemann, Boston
29. Harnden RJ (1989) Technology for Enabling: The Implications for Management Science of a Hermeneutics of Distinction. The University of Aston, Birmingham
30. Jackson MC (1991) Systems Methodology for the Management Sciences. Plenum Press, New York
31. Jackson MC (2000) Systems Approaches to Management. Kluwer Academic/Plenum, New York
32. Kauffman SA (1993) The Origins of Order. Self-Organization and Selection in Evolution. Oxford University Press, New York
33. Klir GJ (1969) An Approach to General Systems Theory. Nostrand, New York
34. Lane DC (1994) Modeling as Learning: A Consultancy Methodology for Enhancing Learning in Management in Management Teams. In: Morecroft J, Sterman JD (eds) Modeling for Learning Organizations. Productivity Press, Portland, pp 205–240
35. Lane DC (1994) With a little help from our friends: How system dynamics and soft OR can learn from each other. Syst Dyn Rev 10(2–3):101–134
36. Lane DC (2006) IFORS' Operational Research Hall of Fame. Jay Wright Forrester. Int Trans Oper Res 13:483–492
37. Lane DC, Oliva R (1998) The Greater Whole: towards a synthesis of system dynamics and soft system methodology. Eur J Oper Res 107(1):214–235
38. La Roche U, Simon M (2000) Geschäftsprozesse simulieren: flexibel und zielorientiert führen mit Fliessmodellen. Orell Füssli, Zürich
39. McCulloch WS (1965) Embodiments of Mind. MIT Press, Cambridge
40. Meadows DH (1980) The Unavoidable A Priori. In: Randers J (ed) Elements of the System Dynamics Method. MIT Press, Cambridge, pp 23–57
41. Meadows DH, Meadows DL, Randers J, Behrens III WW (1972) Limits to Growth. Universe Books, New York
42. Mesarovic MD, Takahara Y (1985) Abstract Systems Theory. Springer, Berlin
43. Midgley G (2000) Systemic Intervention. Philosophy, Methodology, and Practice. Kluwer, New York
44. Miller JG (1978) Living Systems. McGraw-Hill, New York
45. Mingers J (1997) Multi-paradigm Multimethodology. In: Mingers J, Gill A (eds) Multimethodology. Wiley, Chichester
46. Morecroft J (2007) Strategic Modelling and Business Dynamics: a Feedback Systems Approach. Wiley, Chichester
47. Odum HT, Odum EC (2000) Modeling for all Scales: An Introduction to System Simulation. Academic Press, San Diego
48. Pask G (1961) An Approach to Cybernetics. Hutchinson, London
49. Pichler F (1975) Mathematische Systemtheorie. de Gruyter, Berlin
50. Rapoport A (1953) Operational philosophy: Integrating Knowledge and Action. Harper, New York
51. Rapoport A (1986) General System Theory. Essential Concepts and Applications Abacus Press, Turnbridge Wells
52. Richardson GP (1999) Feedback Thought in Social Science and Systems Theory. Pegasus Communications, Waltham (Originally published in 1991)
53. Richmond B (1997) The "Thinking" in systems thinking: How can we make it easier to master? Syst Think 8(2):1–5
54. Sabelli H (2005) Bios: a Study of Creation. World Scientific, Hackensack
55. Schwaninger M (1997) Integrative systems methodology: Heuristic for requisite variety. Int Trans Oper Res 4(4):109–123
56. Schwaninger M (2004) Methodologies in conflict: Achieving synergies between system dynamics and organizational cybernetics. Syst Res Behav Sci 21(4):1–21
57. Schwaninger M (2006) Theories of viability. A comparison. Syst Res Behav Sci 23:337–347
58. Schwaninger M, Groesser S (2008) System dynamics as model-based theory building. Syst Res Behav Sci 25:1–19
59. Schwaninger M, Groesser S (2009) Model Validation: The Quest for Quality in System Dynamics Modeling. Encyclopaedia of Complexity and Systems Science. Springer, New York
60. Schwaninger M, Janovjak M, Ambroz K (2006) Second-order intervention: Enhancing organizational competence and performance. Syst Res Behav Sci 23:529–545
61. Schwaninger M, Pérez Ríos J (2008) System dynamics and cybernetics: A synergetic pair. Syst Dyn Rev 24(2):145–174
62. Senge PM (1990) The Fifth Discipline. The Art and Practice of the Learning Organization. Doubleday, New York
63. Shapiro M, Mandel T, Schwaninger M et al (1996) The Primer Toolbox. International Society for the Systems Sciences, http://www.isss.org/primer/toolbox.htm
64. Simon HA (1969) The Sciences of the Artificial. MIT Press, Cambridge
65. Sterman JD (1994) Learning in and about complex systems. Syst Dyn Rev 10(2–3):291–330
66. Sterman JD (2000) Business Dynamics. Systems Thinking and Modeling for a Complex World. Irwin/McGraw-Hill, Boston
67. Ulrich W (1983) Critical Heuristics of Social Planning. Haupt, Bern
68. Ulrich W (1996) A Primer to Critical Systems Heuristics for Action Researchers. The Centre of Systems Studies, University of Hull, Hull
69. Vennix JAM (1996) Group Model Building. Facilitating Team Learning Using System Dynamics. Wiley, Chichester
70. Vester F, Von Hesler A (1980) Sensitivitätsmodell. Regionale Planungsgemeinschaft Untermain, Frankfurt am Main
71. Von Bertalanffy L (1949) Zu einer allgemeinen Systemlehre. Bl Dtsch Philos 18(3/4) (Excerpts: in Biol Gen 19(1):114–129 and in General System Theory, 1968, Chapter III)
72. Von Bertalanffy L (1950) An outline of general system theory. Br J Philos Sci 1:139–164

73. Von Bertalanffy L (1968) General System Theory. Braziller, New York
74. Von Foerster H (1984) Observing Systems, 2nd edn. Intersystems Publications, Seaside
75. Von Foerster H (ed) (1995) Cybernetics of Cybernetics, 2nd edn. Future Systems, Minneapolis (Originally published in 1974)
76. Von Foerster H, Zopf GW (eds) (1962) Principles of Self-Organization. Pergamon Press, Oxford
77. Vriens D, Achtenbergh J (2006) The social dimension of system dynamics-based modelling. Syst Res Behav Sci 23(4): 553–563
78. Warfield JN (1994) A Science of Generic Design: Managing Complexity through Systems Design, 2nd edn. Iowa State University Press, Ames
79. Warren K (2002) Competitive Strategy Dynamics. Wiley, Chichester
80. Warren K (2008) Strategic Management Dynamics. Wiley, Chichester
81. Wiener N (1948) Cybernetics: Control and Communication in the Animal and in the Machine. MIT Press, Cambridge
82. Wolstenholme E (2003) Towards the definition and use of a core set of archetypal structures in system dynamics. Syst Dyn Rev 19(1):7–26
83. Yolles MA (1998) Cybernetic exploration of methodological complement. Kybern 27(5):527–542

Books and Reviews

Jackson MC (2003) Systems Thinking: Creative Holism for Managers. Wiley, Chichester
Klir GJ (2001) Facets of Systems Science, 2nd edn. Kluwer Academic/Plenum, New York
Midgley G (ed) (2003) Systems Thinking, vol 4. Sage, London
Ragsdell G, Wilby J (eds) (2001) Understanding Complexity. Kluwer Academic/Plenum, New York
Richardson GP (ed) (1996) Modelling for Management. Simulation in Support of Systems Thinking, 2 Volumes. Aldershot, Dartmouth
Schwaninger M (2006) Intelligent Organizations. Powerful Model for Systemic Management. Springer, Berlin
Van Gigch JP (2003) Metadecisions. Rehabilitating Epistemology. Kluwer Academic/Plenum, New York

System Dynamics, Introduction to

Brian Dangerfield
Centre for OR & Applied Statistics, Salford Business School, University of Salford, Salford, UK

When Jay Wright Forrester published his first paper in 1958 he subtitled it *"a major breakthrough for decision-makers"*. At the time some thought this rather an exaggeration if not pompous. Now that 50 years of system dynamics (SD) has elapsed we can at least point to the achievements made and re-state continuing progress in the pages of this section. Was it a 'major breakthrough'? It certainly has the potential to raise the standards in evidence-based policy making to warrant this description and some startlingly good examples of such work will be mentioned here. But after 50 years perhaps one might expect more than has surfaced heretofore.

The key might be connected to the skills required to formulate good SD models – those which address a real-world problem with devastating simplicity and insight. It is deceptively easy to produce an SD model but there are subtleties involved in producing a really effective model for policy purposes. An uplift in modeling skills is something which a subset of the (now significant) amount of published material on SD is aimed at and this section will add to that corpus of work. In addition it will illustrate the extent to which SD applications have spread from its genesis in business to embrace health care, environmental, energy and climate issues, project management, some aspects of biological science and human physiology, governmental and public policy generally, economics (mainly macro), the diffusion of innovations and finally social and economic development. Other applications are being encountered as the power of the methodology is becoming appreciated. It has long since justified the change of title from ***Industrial*** Dynamics (1958) to ***System*** Dynamics (1970 onwards).

Richardson contributes an overview of the basics of SD modeling (see ► System Dynamics, The Basic Elements of). The underlying conceptual framework is that of the information feedback loop together with resource stocks and flows and an endogenous perspective on causation. The simplicity of the loop concept is apt to contribute to the apparent ease with which SD models can be created (along with the icon-based suites of SD software). But the novice reader should appreciate that it can take time to assimilate the modeling skills necessary to execute well an SD model-based application. Practice is essential and the references included will lead to further published material to assist the steep climb up the learning curve. So-called experts are still being confronted with the subtleties of SD modeling after years of involvement.

To place the SD methodology in context, the contribution by Schwaninger (see ► System Dynamics in the Evolution of the Systems Approach) profiles it alongside various others 'systems' based approaches which have emerged in the management and social sciences. Those professing to become experts in SD need to know about the other range of approaches which co-exist in the field of systems science. All these other methodologies have their own enthusiasts and this may even extend to the formation of societies with annual conferences. His Appendix B shows a diagram of the different systems approaches and their interrelationships.

The foundations of the SD methodology can be characterized by certain philosophical issues. Olaya's text (see ► System Dynamics Philosophical Background and Underpinnings) defines a central one as presentationalism, associated with the notion of 'mental models'. A number of other philosophical issues which relate to SD are introduced, including those of positivism and social theory.

The practice of SD when applied to real-world applications essentially involves managerial learning and will often involve an interaction with client teams rather than one individual. How best to organize such structured approaches to participative model building is described by Rouwette and Vennix (see ► Group Model Building). Client participation is required for successful modeling.

If the promotion of learning and understanding is the primary *raison d'etre* of SD, then achievement of this goal in an individual can be a significant accomplishment, especially if that person is the most senior in the client team. But there is a further goal to be pursued should the study fully reap the benefits of the SD methodology: How can we foster *organizational* learning? Maani tackles this head on (see ► System Dynamics and Organizational Learning). He defines the core capabilities of a learning organization and goes on to list the developing literature on organizational learning and, most importantly, how SD can aid the process through learning laboratories and microworlds.

Running an SD model creates a time-path of output behavior covering all the variables it is deemed necessary to include in the model. The various runs of the model are, most frequently, addressed in comparative fashion rather than taken in isolation. They can therefore be described as computer-based scenarios each of which charts a possible but not assured future. Georgantzas (see ► Scenario-Driven Planning with System Dynamics) describes environmental (traditional) scenario generation for which there is a considerable body of literature. But he emphasizes that successful strategy design involves the integration of three things: a knowledge of the business environment; the effects of unstated assumptions about change in the environment and strategy on performance; and finally the need to *compute* the effects on organizational performance. These three facets are accomplished by the process of SD modeling.

Thus far this introductory roadmap has covered all the background for contextualizing and creating an SD model. We now turn to various tasks associated with ex post modeling activities. Three such aspects are covered: model validation; analytical methods to explain behavior and determine dominant loops; and model optimization.

Schwaninger and Groesser (see ► System Dynamics Modeling: Validation for Quality Assurance) range over the various aspects of model validation, beginning with its epistemological foundations. In real-world modeling studies testing and validation is a sine qua non of the process. The range of tests made available and the attention given to the task of validation in the literature mark out SD as unique in the field of management science. Few other methodologies get near to the variety of tests which can be applied to an SD model. The authors consider the range of tests under three headings: model-related context; model structure; and model behavior.

Kampmann and Oliva deal with the behavioral analysis issue (see ► System Dynamics, Analytical Methods for Structural Dominance Analysis in). This activity tries to shed light on the model's dynamic behavior: Why does it behave as it does? What loop structures are responsible for the dominant behavior – and indeed shifts in that behavior where it occurs? In other words, they explore the link between system structure and dynamic behavior. Early methods used eigenvalue analysis but, since then, more sophisticated approaches have been put forward. A major advance will occur when one or more of these is refined enough to be included in an SD software package. This is likely to take some time although an improved user interface showing links glowing with differing degrees of intensity, reflecting their relative importance, is possible in the not-too-distant future.

Dangerfield describes the methods for improving model performance (see ► System Dynamics Models, Optimization of). The task can be categorized under two headings: calibration and policy optimization. The former relates to the determination of optimal parameter sets which deliver the best fit of the model to past time series data. Policy optimization on the other hand seeks to establish policies which deliver the 'best' performance against a suitable metric, such as minimum cost or maximum revenue. Using such an approach can accelerate the learning which comes from repeated runs of the model. Sadly, in the existing SD literature, there is scant evidence of its use in real-world studies.

The methodology of SD exists for no other reason than to offer a quantum leap in the standards of policy analysis. Therefore, any review must include a range through the landscape which defines areas of application. There are eight such areas covered in this section and the choice has been made in the knowledge that there are others which may also have been included and some new areas which are only just being opened up to the tools of SD modeling.

Business Strategy was the genesis of SD applications and rightly takes pride of place. This is the field in which the most numerous SD applications occur. Lyneis (see ► Business Policy and Strategy, System Dynamics Applications to) concentrates on the process of how SD models are used in the task of strategy formulation. He goes on to consider the various drivers of business dynamics such as oscillations in supply chains and boom and bust life cycles. Detailed references are provided for a wide range of business application case studies.

Health care is consuming a higher share of GDP in many Western industrialized countries. This is due to the age profile of the population and advances in pharmacological and medical technologies. It is unsurprising that SD methods have been applied in tackling some of the most high-profile issues in health care and the relatively recent literature is testimony to the success of SD-based analysis. Indeed, it is arguable that some of the best modeling applications have surfaced in this sector. To do justice to the field of health care two contributions were solicited, in part because of the different funding systems which exist on either side of the Atlantic: Wolstenholme surveys the work done by UK and European authors (see ► Health Care in the United Kingdom and Europe, System Dynamics Applications to), whilst Hirsch and Homer concentrate on work published by US authors (see ► Health Care in the United States, System Dynamics Applications to).

Wolstenholme describes work carried out in the UK and Continental Europe but gives particular emphasis to three areas where models have been deployed. He starts with the problem of delayed hospital discharge which generates hospital capacity problems. Epidemiology is also reviewed, in particular research on the epidemiology of HIV/AIDS. Finally, recent work on mental health reform in the UK is described.

Hirsch and Homer note that the system in the USA is comparatively difficult to manage because of its free-market approach and relative lack of regulation. They concentrate on three main areas: disease epidemiology including heart disease and diabetes; substance abuse; and health care capacity and delivery.

Along with health care, the depletion of environmental resources and its effects has consumed many thousands of column inches in printed news media. SD has been employed in the pursuit of more compelling applications in this sector and the efforts go back to the well-known *Limits to Growth* study in 1971–72. Ford charts the most notable efforts which have emerged (see ► System Dynamics Models of Environment, Energy and Climate Change). He ranges over environmental resource problems in the western USA, models for greater understanding of climate change and global warming and concludes with studies in energy, specifically two applications to the electric power industry.

The field of economics is one where SD has received a mostly hostile reception. The statistical economic modeling tool of econometrics has an extensive history and as a preferred modeling methodology seems hard to dislodge. However, there are an increasing number of heterodox economists who are prepared to embrace SD concepts and Radzicki (see ► System Dynamics and Its Contribution to Economics and Economic Modeling) describes the advances taking place. Whilst some of the literature embodies the translation of existing economic models into an SD format (which is a laudable objective) he calls for more economic dynamics models to be built from scratch embodying the best practice in SD modeling. Economic policy is too important to be informed by a single, seemingly unassailable, modeling methodology and it is to be hoped that in the future SD will become even more accepted as a viable tool for use in this field.

In a similar vein comes the contribution of Saeed (see ► Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation). He takes an economic modeling perspective and describes an SD model which explains resource allocation, production and entitlements in a market economy. Its purpose is to understand better how poverty might be reduced in the context of the redistribution of income. A comprehensive listing of the model is provided in an appendix.

The application of SD to public policy generally is dealt with by Andersen, Rich and MacDonald (see ► Public Policy, System Dynamics Applications to). They emphasize how public policy issues are complex, cross organizational boundaries, involve stakeholders with widely different perspectives and evolve over time, such that longer term results may be wholly different from short-term outcomes. Detail is provided for one public policy case involving the Governor's Office of Regulatory Assistance in New York State. They conclude with coverage of studies in a range of public domains such as defense, health care, education and the environment.

One area of SD application has brought the methodology into the legal arena. Disruption and delay in the execution of complex projects invariably finds two parties in dispute. Such disputes often center upon time delays and use of resources on projects – and what might have happened if things had been managed differently. SD models have been employed by parties to such disputes to attempt to justify the occurrence of these events. Howick, Ackermann, Eden and Williams (see ► Delay and Disruption in Complex Projects) report on how cognitive mapping, cause mapping and SD can be fused into what they describe as a cascade model building process. The result is a rigorous process for explaining why a project behaved in a certain way.

New products and processes are emerging at an ever-increasing rate in modern times. We need to understand the myriad mechanisms which are the basis for their rate of adoption. Milling and Maier range over various SD models which have been created to understand and improve the management of the diffusion of innovations (see ▶ Diffusion of Innovations, System Dynamics Analysis of the). From the often-cited Bass diffusion model (1969) the authors develop a series of additional features in a modular fashion. These features include competition, network externalities, dynamic pricing and research and development. They conclude by stressing how it is not possible to offer general recommendations for strategies in dynamic and complex environments; such recommendations can only be given in the context of the specific case under scrutiny.

System Dynamics and Its Contribution to Economics and Economic Modeling

Michael J. Radzicki
Worcester Polytechnic Institute, Worcester, USA

Article Outline

Glossary

Stock Stocks, which are sometimes referred to as "levels" or "states", accumulate (i. e., sum up) the information or material that flows into and out of them. Stocks are thus responsible for decoupling flows, creating delays, preserving system memory, and altering the time shape of flows.

Flow Flows of information or material enter and exit a system's stocks and, in so doing, create a system's dynamics. Stated differently, the net flow into or out of a stock is the stock's rate of change. When human decision making is represented in a system dynamics model, it appears in the system's flow equations. Mathematically, a system's flow equations are ordinary differential equations and their format determines whether or not a system is linear or nonlinear.

Feedback Feedback is the transmission and return of information about the amount of information or material that has accumulated in a system's stocks. When the return of this information reinforces a system's behavior, the loop is said to be positive. Positive loops are responsible for the exponential growth of a system over time. Negative feedback loops represent goal seeking behavior in complex systems. When a negative loop detects a gap between the amount of information or material in a system's stock and the desired amount of information or material, it initiates corrective action. If this corrective action is not significantly delayed, the system will smoothly adjust to its goal. If the corrective action is delayed, however, the system can overshoot or undershoot its goal and the system can oscillate.

Full information maximum likelihood with optimal filtering FIMLOF is a sophisticated technique for estimating the parameters of a system dynamics model, while simultaneously fitting its output to numerical data. Its intellectual origins can be traced to control engineering and the work of Fred Schwepe. David Peterson pioneered a method for adapting FIMLOF for use in system dynamics modeling.

Definition of the Subject

System dynamics is a computer modeling method that has its intellectual origins in control engineering, management science, and digital computing. It was originally created as a tool to help managers better understand and control corporate systems. Today it is applied to problems in a wide variety of academic disciplines, including economics. Of note is that system dynamics models often generate behavior that is both counterintuitive and at odds with traditional economic theory. Historically, this has caused many system dynamics models to be evaluated critically, especially by some economists. However, today economists from several schools of economic thought are beginning to use system dynamics, as they have found it useful for incorporating their nontraditional ideas into formal models.

Introduction

System dynamics is a computer simulation modeling methodology that is used to analyze complex nonlinear dynamic feedback systems for the purposes of generating insight and designing policies that will improve system performance. It was originally created in 1957 by Jay W. Forrester of the Massachusetts Institute of Technology as a method for building computer simulation models of problematic behavior within corporations. The models were used to design and test policies aimed at altering a corporation's structure so that its behavior would improve and become more robust. Today, system dynamics is applied to a large variety of problems in a multitude of academic disciplines, including economics.

System dynamics models are created by identifying and linking the relevant pieces of a system's structure and simulating the behavior generated by that structure. Through an iterative process of structure identification, mapping, and simulation a model emerges that can explain (mimic) a system's problematic behavior and serve as a vehicle for policy design and testing.

From a system dynamics perspective a system's structure consists of stocks, flows, feedback loops, and limiting factors. Stocks can be thought of as bathtubs that accumulate/de-cumulate a system's flows over time. Flows can be thought of as pipe and faucet assemblies that fill or drain the stocks. Mathematically, the process of flows accumulating/de-cumulating in stocks is called integration. The integration process creates all dynamic behavior in the world be it in a physical system, a biological system, or a socioeconomic system. Examples of stocks and flows in economic systems include a stock of inventory and its inflow of production and its outflow of sales, a stock of the book value of a firm's capital and its inflow of investment spending and its outflow of depreciation, and a stock of employed labor and its inflow of hiring and its outflow of labor separations.

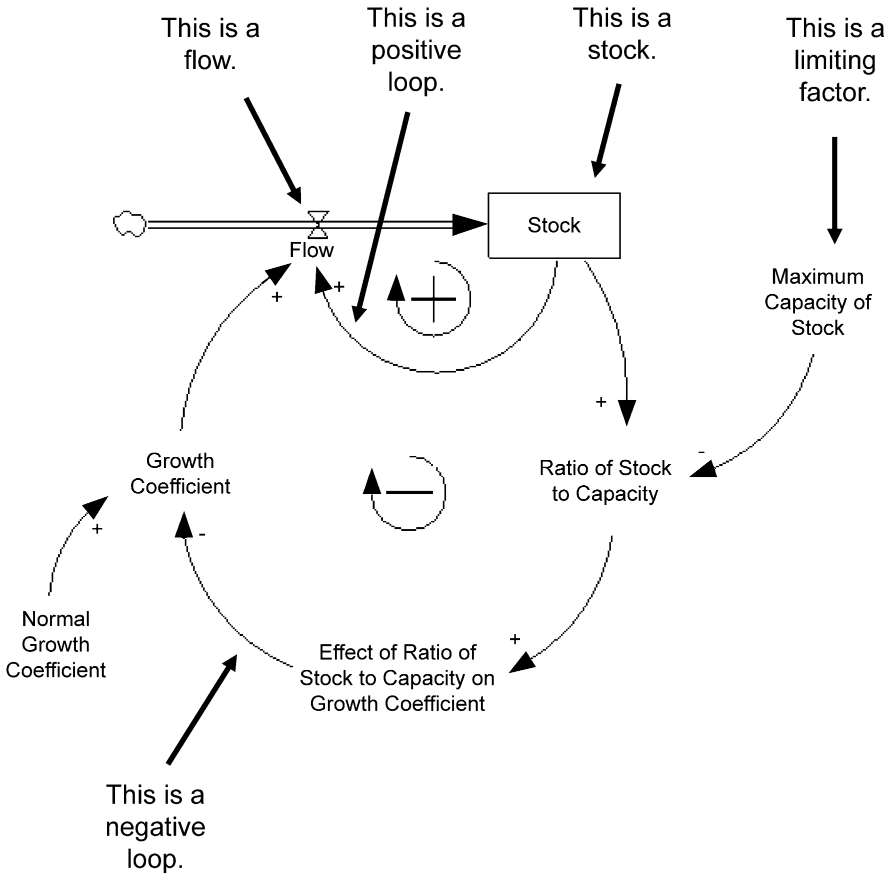
Feedback is the transmission and return of information about the amount of information or material that has accumulated in a system's stocks. Information travels from a stock back to its flow(s) either directly or indirectly, and this movement of information causes the system's faucets to open more, close a bit, close all the way, or stay in the same place. Every feedback loop has to contain at least one stock so that a simultaneous equation situation can be avoided and a model's behavior can be revealed recursively. Loops with a single stock are termed minor, while loops containing more than one stock are termed major.

Two types of feedback loops exist in system dynamics modeling: positive loops and negative loops. Generally speaking, positive loops generate self-reinforcing behavior and are responsible for the growth or decline of a system. Any relationship that can be termed a virtuous or vicious circle is thus a positive feedback loop. Examples of positive loops in economic systems include path dependent processes, increasing returns, speculative bubbles, learning-by-doing, and many of the relationships found in macroeconomic growth theory. Forrester [12], Radzicki and Sterman [46], Moxnes [32], Sterman (Chap. 10 in [55]), Radzicki [44], Ryzhenkov [49], and Weber [58] describe system dynamics models of economic systems that possess dominant positive feedback processes.

Negative feedback loops generate goal-seeking behavior and are responsible for both stabilizing systems and causing them to oscillate. When a negative loop detects a gap between a stock and its goal it initiates corrective action aimed at closing the gap. When this is accomplished without a significant time delay, a system will adjust smoothly to its goal. On the other hand, if there are significant time lags in the corrective actions of a negative loop, it can overshoot or undershoot its goal and cause the system to oscillate. Examples of negative feedback processes in economic systems include equilibrating mechanisms ("auto-pilots") such as simple supply and demand relationships, stock adjustment models for inventory control, any purposeful behavior, and many of the relationships found in macroeconomic business cycle theory. Meadows [27], Mass [26], Low [23], Forrester [12], and Sterman [54] provide examples of system dynamics models that generate cyclical behavior at the macro-economic and micro-economic levels.

From a system dynamics point of view, positive and negative feedback loops fight for control of a system's behavior. The loops that are dominant at any given time determine a system's time path and, if the system is nonlinear, the dominance of the loops can change over time as the system's stocks fill and drain. From this perspective, the dynamic behavior of any economy – that is, the interactions between the trend and the cycle in an economy over time – can be explained as a fight for dominance between the economy's most significant positive and negative feedback loops.

In system dynamics modeling, stocks are usually conceptualized as having limits. That is, stocks are usually seen as being unable to exceed or fall below certain maximum and minimum levels. Indeed, an economic model that can generate, say, either an infinite and/or a negative work-


This is a flow.
This is a positive loop.
This is a stock.
This is a limiting factor.
Stock
Flow
+
+
+
Maximum Capacity of Stock
-
Ratio of Stock to Capacity
Growth Coefficient
-
+
Normal Growth Coefficient
+
Effect of Ratio of Stock to Capacity on Growth Coefficient
This is a negative loop.

System Dynamics and Its Contribution to Economics and Economic Modeling, Figure 1
Simple system dynamics model containing examples of all components of system structure

force would be seen as severely flawed by a system dynamicist. As such, when building a model system dynamicists search for factors that may limit the amount of material or information that the model's stocks can accumulate. Actual socioeconomic systems possess many limiting factors including physical limits (e. g., the number of widgets a machine can produce per unit of time), cognitive limits (e. g., the amount of information an economic agent can remember and act upon), and financial limits (e. g., the maximum balance allowed on a credit card). When limiting factors are included in a system dynamics model, the system's approach to these factors must be described. Generally speaking, this is accomplished with nonlinear relationships. Figure 1 presents a simple system dynamics model that contains examples of all of the components of system structure described above.

Types of Dynamic Simulation

From a system dynamics point of view, solving a dynamic model – any dynamic model – means determining how much material or information has accumulated in each of a system's stocks at every point in time. This can be accomplished in one of two ways – analytically or via simulation. Linear dynamic models can be solved either way. Nonlinear models, except for a few special cases, can only be solved via simulation.

Simulated solutions to dynamic systems can be attained from either a continuous (analog) computer or a discrete (digital) computer. Understanding the basic ideas behind the two approaches is necessary for understanding how economic modeling is undertaken with system dynamics.

In the real world, of course, time unfolds continuously. Yet, devising a way to mimic this process on a machine is a bit tricky. On an analog computer, the continuous flow of economic variables in and out of stocks over time is mimicked by the continuous flow of some physical substance such as electricity or water. A wonderful example of the later case is the Phillips Machine, which simulates an orthodox Keynesian economy (essentially the IS-LM model) with flows of colored water moving through pipes and accumulating in tanks. Barr [2] provides a vivid description of the history and restoration of the Phillips Machine.

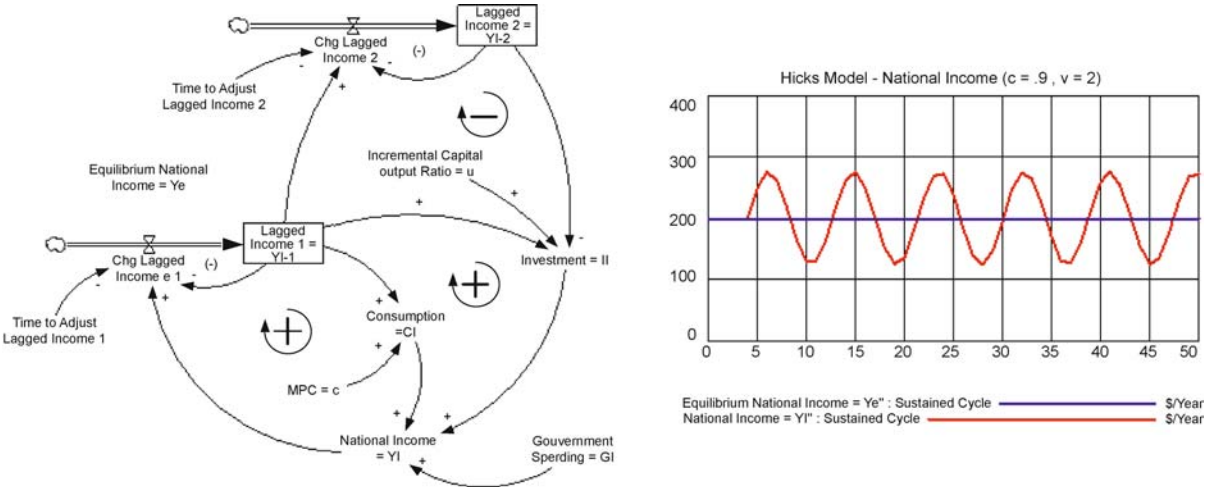
On a digital computer, the continuous flow of economic variables in and out of stocks over time is approximated by specifying the initial amount of material or information in a system's stocks, breaking simulated time into small increments, inching simulated time forward by one of these small increments, calculating the amount of material or information that flowed into and out of the system's stocks during this small interval, and then repeating. The solution to the system will always be approximate because the increment of time cannot be made infinitesimally small and thus simulated time cannot be made perfectly continuous. In fact, on a digital computer a trade-off exists between round-off error and integration error. If the increment of time is made too large, the approximate solution can be poor due to integration error. If the increment of time is made too small, the approximate solution can be ruined due to round-off error.

In system dynamics modeling the "true" behavior of the underlying system is conceptualized to unfold over continuous time. As such, mathematically, a system dynamics model is an ordinary differential equation model. To approximate the solution to a continuous time ordinary differential equation model on a digital (discrete) computer, however, difference equations are used. Unlike traditional difference equation modeling in economics, in which the increment of time is chosen to match economic data (typically a quarter or a year), the increment of time in system dynamics modeling is chosen to yield a solution that is accurate enough for the problem at hand, yet avoids the problems associated with significant round-off and integration error.

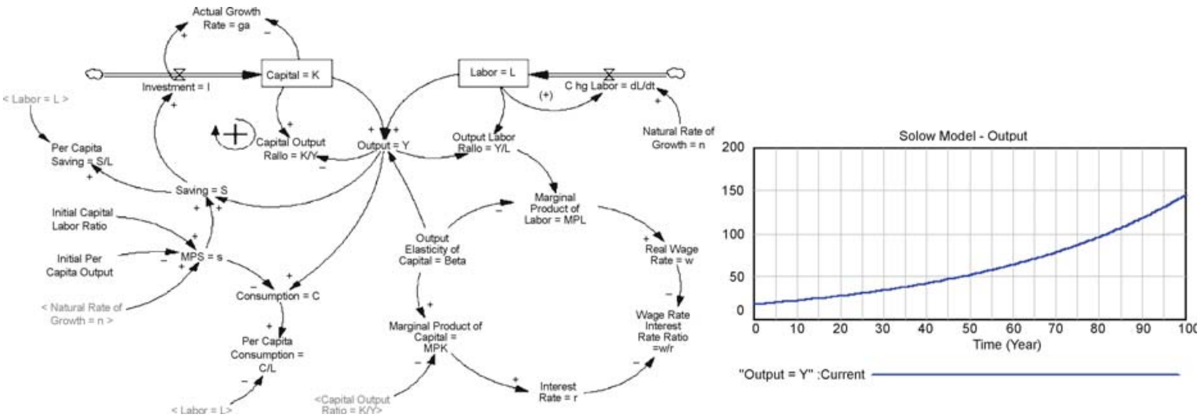
The use of difference equations to approximate the underlying differential equations represented by a system dynamics model provides another interesting option when it comes to economic modeling. Since many well known dynamic economic models have been created with difference equations, they can be recast in a system dynamics format by using the difference equations in the system dynamics software literally as difference equations, and not as a tool to approximate the underlying continuous time system. Although doing this deviates from the original ideas embodied in the system dynamics paradigm, it is occasionally done when a modeler feels that analyzing a difference equation model in a system dynamics format will yield some additional insight.

Translating Existing Economic Models into a System Dynamics Format

There are three principle ways that system dynamics is used for economic modeling. The first involves translating an existing economic model into a system dynamics format, while the second involves creating an economic model from scratch by following the rules and guidelines of the system dynamics paradigm. Forrester [7], Richardson and Pugh [47], Radzicki [42], and Sterman [55] provide extensive details about these rules and guidelines. The former approach is valuable because it enables well-known economic models to be represented in a common format, which makes comparing and contrasting their as-


Lagged Income 2 = YI-2
Chg Lagged Income 2
(-)
Time to Adjust Lagged Income 2
Equilibrium National Income = Ye
Incremental Capital output Ratio = u
Lagged Income 1 = YI-1
Chg Lagged Income e 1
(-)
Time to Adjust Lagged Income 1
Investment = II
Consumption =CI
MPC = c
National Income = YI
Gouvemment Sperding = GI
+
-
Hicks Model - National Income (c = .9 , v = 2)
400
300
200
100
0
0
5
10
15
20
25
30
35
40
45
50
Equilibrium National Income = Ye" : Sustained Cycle
National Income = YI" : Sustained Cycle
$/Year
$/Year

System Dynamics and Its Contribution to Economics and Economic Modeling, Figure 2
System dynamics representation of John Hicks' multiplier-accelerator difference equation model


Actual Growth Rate = ga
Investment = I
Capital = K
Labor = L
C hg Labor = dL/dt
(+)
< Labor = L >
Per Capita Saving = S/L
Capital Output Rallo = K/Y
Output = Y
Output Labor Rallo = Y/L
Natural Rate of Growth = n
Saving = S
Marginal Product of Labor = MPL
Initial Capital Labor Ratio
Output Elasticity of Capital = Beta
Real Wage Rate = w
Initial Per Capita Output
MPS = s
Consumption = C
< Natural Rate of Growth = n >
Wage Rate Interest Rate Ratio =w/r
Per Capita Consumption = C/L
Marginal Product of Capital = MPK
Interest Rate = r
< Labor = L>
<Capital Output Ratio = K/Y>
Solow Model - Output
200
150
100
50
0
0
10
20
30
40
50
60
70
80
90
100
Time (Year)
"Output = Y" :Current

System Dynamics and Its Contribution to Economics and Economic Modeling, Figure 3
System dynamics representation of Robert Solow's ordinary differential equation growth model

sumptions, concepts, structures, behaviors, etc., fairly easy. The latter approach is valuable because it usually yields models that are more realistic and that produce results that are "counterintuitive" [11] and thus thought-provoking.

The third way that system dynamics can be used for economic modeling is a "hybrid" approach in which a well known economic model is translated into a system dynamics format, critiqued, and then improved by modifying it so that it more closely adheres to the principles of system dynamics modeling. This approach attempts to blend the advantages of the first two approaches, although it is more closely related to the former.

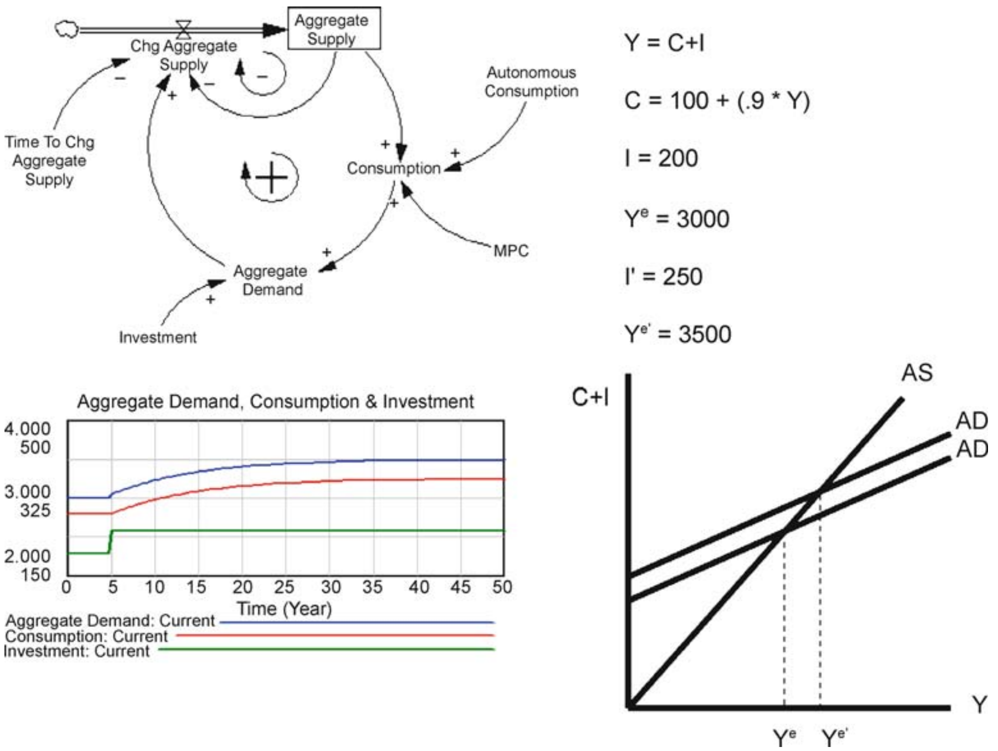
Generally speaking, existing economic models that can be translated into a system dynamics format can be divided into four categories: written, static (mathematical), difference equation, and ordinary differential equation. Existing economic models that have been created in either a difference equation or an ordinary differential equation format can be translated into system dynamics in a fairly straight-forward manner. For example, Fig. 2 presents Sir John Hicks' [21] Multiplier-Accelerator difference equation model in a system dynamics format and Fig. 3 presents the Robert Solow's [52] ordinary differential equation growth model in a system dynamics format.

Translating existing static and written economic models and theories into a system dynamics format is a more formidable task. Written models and theories are often dynamic, yet are described without mathematics. Static models and theories are often presented with mathematics, but lack equations that describe the dynamics of any adjustment processes they may undergo. As such, system dynamicists must devise equations that capture the dynamics being described by the written word or that reveal the adjustment processes that take place when a static system moves from one equilibrium point to another.

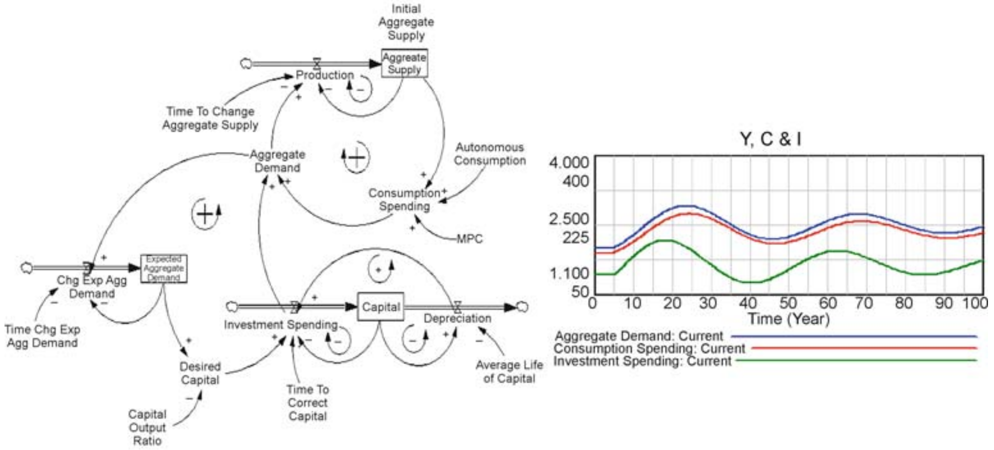
An interesting example of a system dynamics model that was created from a written economic model is Barry Richmond's [48] model of Adam Smith's *Wealth of Nations*. This model was created principally from Robert Heilbronner's [20] written description of Smith's economic system. A classic example of a static model that has been translated into a system dynamics format is a simple two sector Keynesian cross model, as is shown in Fig. 4.

Improving Existing Economic Models with System Dynamics

The simple two sector Keynesian cross model presented in Fig. 4 is an example of a well known economic model that can be improved after it has been translated into a system dynamics format. More specifically, in this example the flow of investment spending in the model does not accumulate anywhere. This violates good system dynamics modeling practice and can be fixed. Figure 5 presents the improved version of the Keynesian Cross model, which now more closely adheres to the system dynamics paradigm. Other well known examples of classic economics models that have been improved after they have been translated into a system dynamics format and made to conform more closely with good system dynamics modeling practice include the cobweb model [27], Sir John Hicks' multiplier-accelerator model [23], the IS-LM/AD-AS model [13,59], Dale Jorgenson's investment model [51], William Nordhaus' [34] DICE climate change model [4,5], and basic micro economic supply and demand mechanisms [24]. Low's improvement of Hicks' model is particularly interesting because it results



System Dynamics and Its Contribution to Economics and Economic Modeling, Figure 4
Simple two sector Keynesian cross model in a system dynamics format



System Dynamics and Its Contribution to Economics and Economic Modeling, Figure 5
Improved simple two sector Keynesian cross model

in a model that closely resembles Bill Phillips' [40] multiplier-accelerator model. Senge and Fiddaman's contributions are also very interesting because they demonstrate how the original economic models are special cases of their more general system dynamics formulations.

Creating Economic Dynamics Models from Scratch

Although translating well known economic models into a system dynamics format can arguably make them easier to understand and use, system dynamicists believe that the "proper" way to model an economic system that is experiencing a problem is to do so from scratch while following good system dynamics modeling practice. Unlike orthodox economists who generally follow a deductive, logical positivist approach to modeling, system dynamicists follow an inductive pattern modeling or case study process. More specifically, a system dynamicist approaches an economic problem like a detective who is iteratively piecing together an explanation at a crime scene. All types of data that are deemed relevant to the problem are considered including numerical, written, and mental information. The system dynamicist is guided in the pattern modeling process by the perceived facts of the case, as well as by real typologies (termed "generic structures" in system dynamics) and principles of systems. Real typologies are commonalities that have been found to exist in different pattern models and principles of systems are commonalities that have been found to exist in different real typologies. Paich [36] discusses generic structures at length and Forrester [8] lays out a set of principles of systems.

Examples of a real typologies in economics include Forrester's [9] *Urban Dynamics* model, which can reproduce the behavior of many different cities when properly parametrized for those cities, and Homer's [22] model of the diffusion of new medical technologies into the market place, which can explain the behavior of a wide variety of medical technologies when properly parametrized for those technologies. Examples of fundamental principles of systems include the principle of accumulation, which states that the dynamic behavior of any system is due to flows accumulating in stocks, and the notion of stocks and flows being components of feedback loops. The parallels for these principles in economics can be found in modern Post Keynesian economics, in which modelers try to build "stock-flow consistent models," and in institutional economics, in which the principle of "circular and cumulative causation" is deemed to be a fundamental cause of economic dynamics. Radzicki [41,43,45] lays out the case for the parallels that exist between methodological concepts in system dynamics and methodological concepts in various schools of economic thought.

The economic models that have been historically created from scratch by following the system dynamics paradigm have tended to be fairly large in scale. Forrester's [12] national economic model is a classic example, as are the macroeconomic models created by Sterman [53], the Millennium Institute [31], Radzicki [45], Wheat [59], and Yamaguchi [60]. Dangerfield [3] has developed a model of Sarawak (E. Malaysia) to analyze and plan for economic transition from a production economy to a knowledge-based one. With the exception of Radzicki [45], whose model is based on ideas from Post Keynesian and institutional economics, these models, by and large, embody orthodox economic relationships.

Model Validity

When a system dynamics model of an economic system that is experiencing a problem is built from scratch, the modeling process is typically quite different from that which is undertaken in traditional economics. As such, the question is raised as to whether or not an original system dynamics model is in any sense "valid".

System dynamicists follow a "pattern modeling" approach [41] and do not believe that models should be judged in a binary fashion as either "valid" or "invalid". Rather, they argue that confidence in models can be generated along multiple dimensions. More specifically, system dynamicists such as Peterson [38], Forrester and Senge [16] and Barlas [1] have developed a comprehensive series of tests that can be applied to a model's structure and behavior and they argue that the more tests a model can pass, the more confidence a model builder or user should place in its results. Even more fundamentally, however, Forrester [13] has argued that the real value generated through the use of system dynamics comes, not from any particular model, but from the modeling *process* itself. In other words, it is through the iterative *process* of model conceptualization, creation, simulation, and revision that true learning and insight are generated, and *not* through interaction with the resulting model.

Another issue that lies under the umbrella of model validity involves fitting models to time series data so that parameters can be estimated and confidence in model results can be raised. In orthodox economics, of course, econometric modeling is almost universally employed when doing empirical research. Orthodox economic theory dictates the structure of the econometric model and powerful statistical techniques are used to tease out parameter values from numerical data.

System dynamicists, on the other hand, have traditionally argued that it is not necessary to tightly fit models to time series data for the purposes of parameter estimation and confidence building. This is because:

1. the battery of tests that are used to build confidence in system dynamics models go well beyond basic econometric analysis;
2. the particular (measured) time path that an actual economic system happened to take is merely one of an infinite number of paths that it could have taken and is a result of the particular stream of random shocks that

happened to be historically processed by its structure. As such, it is more important for a model to mimic the basic character of the data, rather than fit it point-by-point [14];

3. utilizing the pattern modeling/case study approach enables the modeler to obtain parameter values via observation below the level of aggregation in the model, rather than via statistical analysis [18];
4. the result of a system dynamics modeling intervention is typically a set of policies that improve system performance and increase system robustness. Such policies are usually feedback-based rules (i. e., changes to institutional structure) that do not require the accurate point prediction of system variables.

Although the arguments against the need to fit models to time series data are well known in system dynamics, many system dynamicists feel that it is still a worthwhile activity because it adds credibility to a modeling study. Moreover, in modern times, advances in software technology have made this process relatively easy and inexpensive. Although several techniques for estimating the parameters of a system dynamics model from numerical data have been devised, perhaps the most interesting is David Peterson's [38,39] Full Information Maximum Likelihood with Optimal Filtering (FIMLOF). Figure 5 presents a run from the Harrod growth model, to which an adaptive expectations structure has been added, after it has been fit via FIMLOF to real GDP and labor supply data for the United States economy for the years 1929–2002. The fit is excellent and the estimated parameter values are consistent with those from more traditional econometric studies. See Radzicki [44] for a detailed description of the model and its parameter estimates.

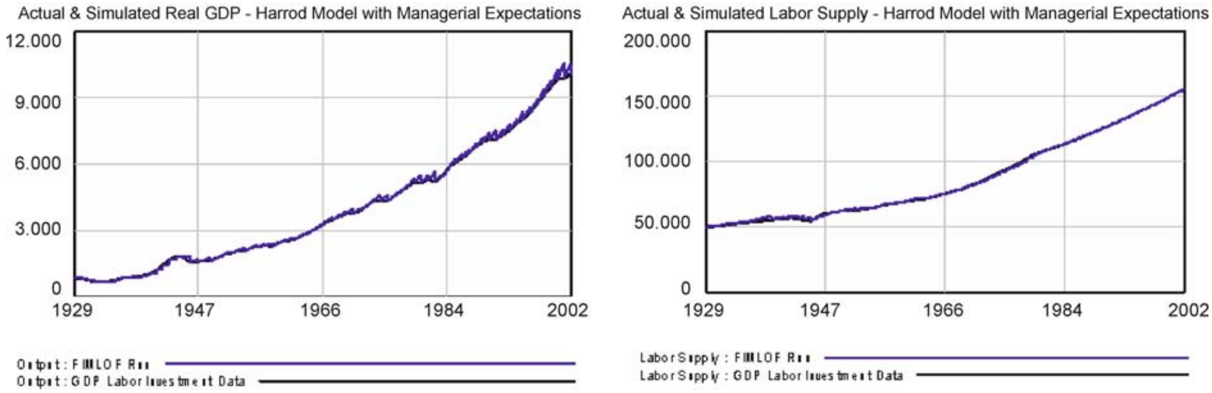
Controversies

Since system dynamics modeling is undertaken in a way that is significantly different from traditional economic modeling, it should come as no surprise that many economists have been extremely critical of some system dynamics models of economic systems. For example, Forrester's [9] *Urban Dynamics* and [10] *World Dynamics* models have come under severe attack by economists, as has (to a lesser degree) his national economic model. On the other hand, the first paper in the field of system dynamics is Forrester [6], which is essentially a critique of traditional economic modeling.

Greenberger et al. [19] present a nice overview of the controversies surrounding the *Urban Dynamics* and *World Dynamics* models. Forrester and his colleagues' replies to criticisms of the *Urban Dynamics* model are contained in Mass [25] and Schroeder et al. [50].

One of the harshest critics of the *World Dynamics* (WORLD2) model has been Nordhaus [33]. Nordhaus [35] has also very critical of the well known follow-up study to *World Dynamics* known as *The Limits to Growth* [28]. Meadows et al. [29,30] contain updates to the original *Limits to Growth* (WORLD3) model, as well as replies to the world modeling critics.

Forrester [12] presents a nice overview of his national economic model, and the critiques by Stolwijk [57] and Zellner [61] are typical of the attitude of the professional economists toward macroeconomic modeling that is undertaken by following the traditional system dynamics paradigm. The criticism of Forrester's national economic model by the economics profession has probably been less severe, relative to the criticisms of the *Urban Dynamics* and world models, because most of its details are still largely unpublished at the time of this writing.



System Dynamics and Its Contribution to Economics and Economic Modeling, Figure 6
Fit of the Harrod growth model to US macroeconomic data for the years 1929–2002

Another interesting and timely example of the sort of controversy surrounding system dynamics modeling in economics is provided by Sterman and Richardson [56]. In this paper they present a technique for testing whether Hubbert's lifecycle method or the geologic analogy method yields superior estimates of the ultimately recoverable amount of petroleum resources. This study was motivated by a disagreement with a traditionally trained economist over the proper way to conceptualize this issue. Sterman and Richardson devised a clever synthetic data experiment in which a system dynamics model serves as the "real world" with a known ultimately recoverable amount of oil. Hubbert's method and the geologic analogy method are then programmed into the model so they can "watch" the data being generated by the "real world" and provide dynamic estimates of the "known" ultimately recoverable stock of oil. The results showed that Hubbert's method was quite accurate, although it had a tendency to somewhat underestimate the ultimately recoverable amount of oil, while the geologic analogy method tended to overshoot the resource base quite substantially.

Future Directions

Historically, system dynamicists who have engaged in economic modeling have almost never been trained as professional economists. As such, they have had the advantage of being able to think about economic problems differently from those who have been trained along traditional lines, but have also suffered the cost of being seen as "amateurs" or "boy economists" [41] by members of the economics profession. The good news is that there are currently several schools of economic thought, populated by professional economists, in which system dynamics fits quite harmoniously. These include Post Keynesian economics, institutional economics, ecological economics, and behavioral economics. Historically, the economists in these schools have rejected many of the tenets of traditional economics, including most of its formal modeling methods, yet have failed to embrace alternative modeling techniques because they were all seen as inadequate for representing the concepts they felt were important. However in the modern era, with computers having become ubiquitous and simulation having become in some sense routine, system dynamics is increasingly being accepted as an appropriate tool for use in these schools of economic thought. The future of economics and system dynamics will most probably be defined by the economists who work within these schools of thought, as well as by their students. The diffusion of system dynamics models of economic systems through their translation into user-friendly interactive "learning environments" that are available over the world wide web will most likely also be of great importance (see [24,59]).

Bibliography

Primary Literature

1. Barlas Y (1989) Multiple Tests for Validation of System Dynamics Type of Simulation Models. Eur J Operat Res 42(1):59–87
2. Barr N (1988) The Phillips Machine. LSE Q Winter 2(4):305–337
3. Dangerfield BC (2007) System dynamics advances strategic economic transition planning in a developing nation. In: Qudrat-Ullah H, Spector M, Davidsen P (eds) Complex decision-making: Theory & practice. Springer, New York, pp 185–209
4. Fiddaman T (1997) Feedback complexity in integrated climate-economy models. Ph D Dissertation, Sloan School of Management, Massachusetts Institute of Technology. Available from http://www.systemdynamics.org/
5. Fiddaman T (2002) Exploring policy options with a behavioral climate-economy model. Syst Dyn Rev 18(2):243–267
6. Forrester J (1957) Dynamic models of economic systems and industrial organizations. System Dynamics Group Memo D-0. Massachusetts Institute of Technology. Available from http://www.systemdynamics.org/
7. Forrester J (1961) Industrial dynamics. Pegasus Communications, Inc., Waltham
8. Forrester J (1968) Principles of systems. MIT Press, Cambridge
9. Forrester J (1969) Urban dynamics. Pegasus Communications, Inc., Waltham
10. Forrester J (1971) World dynamics. Pegasus Communications, Inc., Waltham
11. Forrester J (1975) Counterintuitive behavior of social systems. In: Forrester J (ed) Collected papers of Jay W Forrester, Pegasus Communications, Inc., Waltham, pp 211–244
12. Forrester J (1980) Information sources for modeling the national economy. J Am Statist Assoc 75(371):555–567
13. Forrester J (1985) 'The' model versus a modeling 'process'. Syst Dyn Rev 1(1 and 2):133–134
14. Forrester J (2003) Economic theory for the new millennium. In: Eberlein R, Diker V, Langer R, Rowe J (eds) Proceedings of the Twenty-First Annual Conference of the System Dynamics Society. Available from http://www.systemdynamics.org/
15. Forrester J, Low G, Mass N (1974) The debate on world dynamics: A response to Nordhaus. Policy Sci 5:169–190
16. Forrester J, Senge P (1980) Tests for building confidence in system dynamics models. In: Legasto Jr AA, Forrester JW, Lyneis JM (eds) TIMS Studies in the Management Sciences: System Dynamics, vol 14. North Holland Publishing Company, Amsterdam, pp 209–228
17. Forrester N (1982) A dynamic synthesis of basic macroeconomic theory: Implications for stabilization policy analysis, Ph D Dissertation, Alfred P Sloan School of Management, Massachusetts Institute of Technology. Available from http://www.systemdynamics.org/
18. Graham A (1980) Parameter estimation in system dynamics modeling. In: Randers J (ed) Elements of the system dynamics method. Pegasus Communications, Inc., Waltham, pp 143–161
19. Greenberger M, Crenson M, Crissey B (1976) Models in the policy process: Public decision making in the computer era, Russell Sage Foundation, New York

20. Heilbroner R (1980) The worldly philosophers, 5th edn. Simon & Schuster, New York
21. Hicks J (1950) A contribution to the theory of the trade cycle. Oxford University Press, London
22. Homer J (1987) A diffusion model with application to evolving medical technologies. Technol Forecast Soc Change 31(3):197–218
23. Low G (1980) The multiplier-accelerator model of business cycles interpreted from a system dynamics perspective. In: Randers J (ed) Elements of the system dynamics method. Pegasus Communications, Inc., Waltham, pp 76–94
24. Mashayekhi A, Vakili K, Foroughi H, Hadavandi M (2006) Supply demand world: An interactive learning environment for teaching microeconomics. In: Grosler A, Rouwette A, Langer R, Rowe J, Yanni J (eds) Proceedings of the of the Twenty-Fourth International Conference of the System Dynamics Society. Available at http://www.systemdynamics.org/
25. Mass N (1974) Readings in urban dynamics, vol I. Wright-Allen Press, Cambridge
26. Mass N (1975) Economic cycles: An analysis of the underlying causes. MIT Press, Cambridge
27. Meadows D (1970) Dynamics of commodity production cycles. Massachusetts MIT Press, Cambridge
28. Meadows D, Meadows D, Randers J, Behrens III W (1972) The limits to growth: A report for the Club of Rome's project on the predicament of mankind. Universe Books, New York
29. Meadows D, Meadows D, Randers J (1992) Beyond the limits: Confronting global collapse, envisioning a sustainable future. Chelsea Green Publishing, White River Junction
30. Meadows D, Meadows D, Randers J (2002) Limits to growth: The 30-year update. Chelsea Green Publishing, White River Junction
31. Millennium Institute (2007) Introduction and Purpose of Threshold 21, http://www.millennium-institute.org/resources/elibrary/papers/T21Overview.pdf
32. Moxnes E (1992) Positive feedback economics and the competition between 'hard' and 'soft' energy supplies. J Sci Ind Res 51(March):257–265
33. Nordhaus W (1972) World dynamics: Measurement without data. Econom J 83(332):1156–1183
34. Nordhaus W (1992) The "DICE" model: Background and structure of a dynamic integrated climate-economy model of the economics of global warming. Cowles Foundation for Research in Economics at Yale University, Discussion Paper No. 1009
35. Nordhaus W (1992) Lethal model 2: The limits to growth revisited. Brookings Pap Econo Activity 2:1–59
36. Paich M (1985) Generic structures. Syst Dyn Rev 1(1 and 2):126–132
37. Paich M (1994) Managing the global commons. MIT Press, Cambridge
38. Peterson D (1975) Hypothesis, estimation, and validation of dynamic social models – energy demand modeling. Ph D Dissertation, Department of Electrical Engineering, Massachusetts Institute of Technology. Available from http://www.systemdynamics.org/
39. Peterson D (1980) Statistical tools for system dynamics. In: Randers J (ed) Elements of the system dynamics method. Pegasus Communications, Inc., Waltham, pp 226–245
40. Phillips W (1954) Stabilization policy in a closed economy. Econom J 64(254):290–323
41. Radzicki M (1990) Methodologia oeconomiae et systematis dynamis. Syst Dyn Rev 6(2):123–147
42. Radzicki M (1997) Introduction to system dynamics. Free web-based system dynamics tutorial. Available at http://www.systemdynamics.org/DL-IntroSysDyn/index.html
43. Radzicki M (2003) Mr. Hamilton, Mr. Forrester and a foundation for evolutionary economics. J Econom Issues 37(1):133–173
44. Radzicki M (2004) Expectation formation and parameter estimation in nonergodic systems: the system dynamics approach to post Keynesian-institutional economics. In: Kennedy M, Winch G, Langer R, Rowe J, Yanni J (eds) Proceedings of the Twenty-Second International Conference of the System Dynamics Society. Available at http://www.systemdynamics.org/
45. Radzicki M (2007) Institutional economics, post keynesian economics, and system dynamics: Three strands of a heterodox economics braid. In: Harvey JT, Garnett Jr. RF (eds) The future of heterodox economics. University of Michigan Press, Ann Arbor
46. Radzicki M, Sterman J (1994) Evolutionary economics and system dynamics. In: Englund R (ed) Evolutionary concepts in contemporary economics. University of Michigan Press, Ann Arbor, pp 61–89
47. Richardson G, Pugh A (1981) Introduction to system dynamics modeling with DYNAMO. Pegasus Communications, Inc., Waltham
48. Richmond B (1985) Conversing with a classic thinker: An illustration from economics. Users Guide to STELLA, Chapt 7. High Performance Systems, Inc., Lyme, New Hampshire, pp 75–94
49. Ryzhenkov A (2007) Controlling employment, profitability and proved non-renewable reserves in a theoretical model of the US economy. In: Sterman J, Oliva R, Langer R, Rowe J, Yanni J (eds) Proceedings of the of the Twenty-Fifth International Conference of the System Dynamics Society. Available at http://www.systemdynamics.org/
50. Schroeder W, Sweeney R, Alfeld L (1975) Readings in Urban Dynamics, vol II. Wright-Allen Press, Cambridge
51. Senge P (1980) A system dynamics approach to investment function formulation and testing. Soc-Econ Plan Sci 14:269–280
52. Solow R (1956) A contribution to the theory of economic growth. Q J Econom 70:65–94
53. Sterman J (1981) The energy transition and the economy: A system dynamics approach. Ph D Dissertation, Alfred P Sloan School of Management, Massachusetts Institute of Technology. Available at http://www.systemdynamics.org/
54. Sterman J (1985) A behavioral model of the economic long wave. J Econom Behav Organ 6:17–53
55. Sterman J (2000) Business dynamics: Systems thinking and modeling for a complex world. Irwin-McGraw-Hill, New York
56. Sterman J, Richardson G (1985) An experiment to evaluate methods for estimating fossil fuel resources. J Forecast 4(2):197–226
57. Stolwijk J (1980) Comment on 'information sources for modeling the national economy' by Jay W Forrester. J Am Stat Assoc 75(371):569–572
58. Weber L (2007) Understanding recent developments in growth theory. In: Sterman J, Oliva R, Langer R, Rowe J, Yanni J (eds) Proceedings of the of the Twenty-Fifth International Conference of the System Dynamics Society. Available at http://www.systemdynamics.org/

59. Wheat D (2007) The feedback method of teaching macroeconomics: Is it effective? Syst Dyn Rev 23(4):391–413
60. Yamaguchi K (2007) Balance of payments and foreign exchange dynamics – S D Macroeconomic Modeling (4). In: Sterman J, Oliva R, Langer R, Rowe J, Yanni J (eds) Proceedings of the of the Twenty-Fifth International Conference of the System Dynamics Society. Available at http://www.systemdynamics.org/
61. Zellner A(1980) Comment on 'information sources for modeling the national economy' by Jay W Forrester. J Am Stat Assoc 75(371):567–569

Books and Reviews

Alfeld L, Graham A (1976) Introduction to urban dynamics. Wright-Allen Press, Cambridge

Lyneis J (1980) Corporate planning and policy design: A system dynamics approach. PA Consulting, Cambridge

Meadows D, Meadows D (1973) Toward global equilibrium: Collected papers. Wright-Allen Press, Cambridge

Meadows D, Behrens III W, Meadows D, Naill R, Randers J, Zahn E (1974) Dynamics of growth in a finite world. Wright-Allen Press, Cambridge

Meadows D, Robinson J (1985) The electronic oracle: Computer models and social decisions. Wiley, New York

Richardson G (1999) Feedback thought in social science and systems theory. Pegasus Communications, Inc., Waltham

Sterman J (1988) A skeptic's guide to computer models. In: Grant L (ed) Foresight and National Decisions. University Press of America, Lanham, pp 133–169, Revised and Reprinted in: Barney G, Kreutzer W, Garrett M (1991) Managing a Nation. Westview Press, Boulder, pp 209–230

System Dynamics Modeling: Validation for Quality Assurance

Markus Schwaninger, Stefan Groesser
Institute of Management, University of St. Gallen,
St. Gallen, Switzerland

Article Outline

Glossary

Model/model system A model is a simplified representation of a real system. Models can be descriptive or prescriptive (normative). Their functions can be to enable explanation, anticipation or design. A distinction used in this contribution is between causal and non-causal models, with System Dynamics models being of the former type. The term *model system* is used to stress the systemic character of a model; this serves to identify it as an organized whole of variables and relationships on the one hand, and to distinguish it from the *real system* which is to be modeled, on the other.

Model validity A model's property of adequately reflecting the system modeled. Validity is the primary measure of model quality. It is a matter of degree, not a dichotomized property.

Model purpose The goal for which a model is designed or the function it is intended to fulfill. The model purpose is closely linked to the end-model user or model owner. Model purpose is the criterion for the choice of a model's boundary and design.

Modeling process The process involving phases such as problem articulation, boundary selection, development of a dynamic hypothesis, model formulation, model testing, policy formulation and policy evaluation [28]. The modeling process is followed by model use and implementation, i. e., the realization of actions designed or facilitated by the use of the model.

Validation process Validation is the process by which model validity is enhanced systematically. It consists in gradually building confidence in the usefulness of a model by applying validation tests as outlined in this chapter. In principle, validation pervades all phases of the modeling process, and, in addition, extends into the phases of model use and implementation.

Definition of the Subject

The present chapter addresses the question of building better models. This is crucial for coping with complexity in general, and in particular for the management of dynamic systems. Both the epistemological and the methodological-technological aspects of model validation for the achievement of high-quality models are discussed. The focus is on formal models, i. e. those formulated in a stringent, logical, and mostly mathematical language.

Introduction

The etymological root of *valid* is the Latin word *validus*, which denotes attributes such as strong, powerful and firm. A valid model, then, is well-founded and difficult to reject because it accurately represents the perceived real system which it is supposed to reflect. This system can be either one that already exists or one that is being constructed, or even anticipated, by a modeler or a group of modelers.

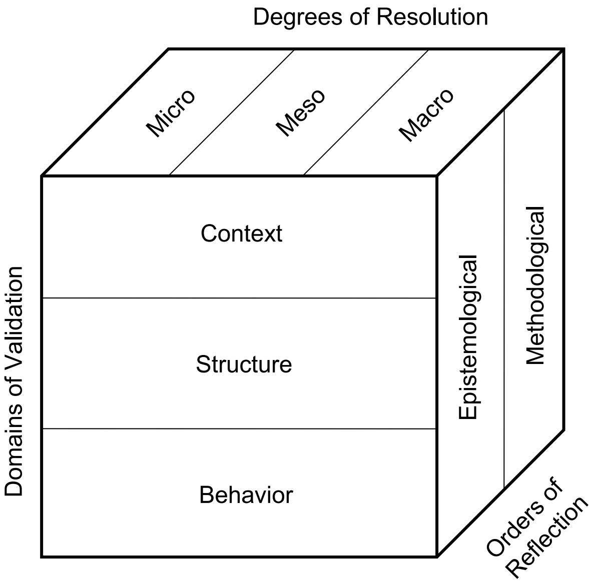
Validation standards in System Dynamics are more rigorous than those of many other methodologies. Let us distinguish between two types of mathematical models, which are fundamentally different: Causal, theory-like models and non-causal, statistical (correlational) models [4]. The former are explanatory, i. e., they embody theory about the functioning of a real system. The latter are descriptive and express observed associations among different elements of a real system. System Dynamics models are causal models.

Non-causal models are tested globally, in that the statistical fit between model and data series from the real system under study is assessed. If the fit is satisfactory, the model is considered to be accurate ("valid", "true"). In contrast, system dynamicists postulate that models be not only right, but right for the right reasons. As the models are made up of causal interdependencies, accuracy is required for each and every variable and relationship. The following principle applies: if only one component of the model is shown to be wrong, the whole model is rejected even if the overall model output fits the data [4]. This strict standard is conducive to high-quality modeling practice.

A model is an abstract version of a perceived reality. Simulation is a way of experimenting with mathematical models to gain insights and to employ these to improve the real system under study. It is often said that System Dynamics models should portray problems or issues, not systems. This statement must be interpreted in the sense that one should not try to set the boundaries of the model too widely, but rather give the model a focus by concentrating on an object in accordance with the specific purpose of the model. In a narrower definition, even an issue or problem can be conceived of as a "system", i. e., "a portion of the world sufficiently well defined to be the subject of study" [21]. Validity then consists in a stringent correspondence between model system and real system.

We will treat the issue of model validation as a means of assuring high-quality models. We interject that validity is not the only criterion of model quality, other criteria including parsimony, ease-of-use, practicality, importance, etc. [22].

In the following, the epistemological foundations of model validity are reviewed (Sect. "Epistemological Foundations"). Then, an overview of the methods for assuring model validity is given (Sect. "Validation Methods"). Further, the survey includes an overview of the validation process (Sect. "Validation Process") and our final conclusions (Sect. "Synopsis and Outlook").



System Dynamics Modeling: Validation for Quality Assurance, Figure 1

The Validation Cube – A frame of reference showing three dimensions of the validation topic

The substance of this article will be made more palpable by means of the following frame of reference. We call it the Validation Cube. The diagram in Fig. 1 shows three dimensions of the validation topic:

- *Orders of Reflection:* We distinguish between an *epistemological* and a *methodological* layer. These define the objects of the next two Sects. "Epistemological Foundations" and "Validation Methods".
- *Domains of Validation:* The three domains, *context*, *structure* and *behavior* refer to the groups of validation methods as described in Sect. "Validation Methods".
- *Degrees of Resolution:* We address the different granularities of models. *Micro* refers to the smallest building blocks of models (e. g., variables or small sets of variables), *meso* to modules which constitute a model, and *macro* to the model as a whole.

Epistemological Foundations

Epistemology is the theory that enquires into the nature and grounds of knowledge: "What can we know and how do we know it?" [13]. These questions are of utmost importance when dealing with models and their validity, because a method of validation is only as good as its epistemological basis.

We can only briefly refer to the antecedents of the epistemological perspective inherent in the idea of model validation as commonly held today in the community of system dynamicists. One could go back to Socrates who, in Plato's *Republic* (fourth century BC), addressed the problematic relationship between reality, image and knowledge. One could also refer to John Locke (seventeenth century), the first British empiricist who maintained that ideas could come only from experience, while admitting that our knowledge about external objects is uncertain. We will address the philosophical movements of the nineteenth and twentieth centuries, which are direct sources of the epistemology which is important for model validation. The reader may kindly excuse us for certain massive simplifications that we are obliged to make.

What will be said here about theories applies equally to formal models. In System Dynamics, models either embody theories or they are considered essential components of theories. In addition, processes of modeling and theory-building are of the same nature; a model, like any theory, is built and improved in a dialectic of propositions and refutations [22].

Positivism and Critique

Positivism is a scientific doctrine founded by Auguste Comte (nineteenth century) which raises the *positive* to the principle of all scientific knowledge. "Positive", in this context, is not meant to be the opposite of negative, but the given, factual, or indubitably existent. The positive is associated with features such as being real, useful, certain, and precise. Positivism confines science to the observable and manipulable, drawing on the mathematical, empirical orientation of the natural sciences as its paragon. The objectivist claim of positivism is that things exist independently of the mind and that truths are detached from human values and beliefs. This stance calls for models that approximate an objective reality.

A younger development in this vein is the school of logical positivism, also logical empiricism (with Schlick, Neurath, Hempel, etc.), which concentrates on the problem of meaning and has developed the verifiability principle: Something is meaningful only if verifiable empirically, i. e., ultimately by observation through the senses. To verify here means to show to be true [13]. For the logical positivists, the method of verification is the essence of theory-building. Tests of theories hinge on their confirmation by facts. In System Dynamics, testing models on real-world data is a core component of validation.

Positivism has been criticized for being reductionist, i. e., for its tendency to reduce concepts to simpler or empirically more accessible ones, and to conceive of learning as an accumulation of particular details. The critique has also asserted that there is no theory-independent identification of facts, and therefore different theories cannot be tested by means of the same data [6]. Another objection maintains that social facts are not merely given, but produced by human action, and that they are subject to interpretation [23]. These arguments introduce the principle of relativity, which is of crucial importance for the field of model validation: A model is a subjective construction by an observer.

Pragmatism – A Challenge to Positivism

Pragmatism, which arose in the second half of the nineteenth century, emphasizes action and the practical consequences of thinking. Its founder, Charles Sanders Peirce, was interested in the effects that the meaning of scientific concepts could have on human experience and action. He defined truth as "the opinion which is fated to be ultimately agreed to by all who investigate" [13], whereby truth is linked to consensual validation. For pragmatists, truth is in what works (Ferdinand Schiller) or satisfies us (John Dewey), and what we find believable and consistent: " 'The true' … is only the expedient in the way of our thinking", and "truth is *made* … in the course of experience." (see p. 581 and p. 583 in [11]).

Pragmatism is often erroneously disdained for supposedly being a crass variety of utilitarianism and embodying a crude instrumentalist rationality. A more accurate view considers the fact that pragmatists are not satisfied with a mere ascertainment of truth; instead they ask: "If an idea or assumption is true, does this make a concrete difference to the life of people? How can this truth be actualized?" In other words, pragmatism does not crudely equate truth and utility. It rather postulates that those truths which are useful to people ought to be put into practice [23].

Pragmatism introduces the criteria of confidence and usefulness, which are more operational as guides to the evaluation of experiments than is the notion of an absolute truth, which is unattainable in the realm of human affairs. At the same time, pragmatism triggers a crucial insight for the context of model-building: The validity of a model depends not only on the absolute quality of that model but also hinges on its suitability with respect to a purpose [7]. In the context of model validation, then, truth is a relative property; more exactly, a *truth* holds for a limited domain only.

More Challenges to Positivism

We discuss three more challenges to positivism in the twentieth century. First, Thomas Kuhn's theory of scientific revolutions [12]: Kuhn shows, by means of historical cases, that in the sphere of science, generally accepted ways of looking at the world ("paradigms") change over time through fundamental shifts. Therefore, the activities of a scientist are largely shaped by the dominant scientific worldview. Second, Willard Van Orman Quine and Wilfrid Sellars argue that knowledge creation and theory-building is a holistic, conversational process, as opposed to the reductionist and confrontational views [4].

Both of these movements contribute to our understanding of how real systems are to be modeled and validated: as organized wholes, and consciously with respect to the values and beliefs underlying a given modeling process. This approach adheres to the spirit of models themselves, by means of which the behavior of whole systems can be simulated and tested on their inherent assumptions.

A third challenge is presented by the interpretive streams of epistemology (for an overview, see [9]). Among them, a main force which expands the possibilities of scientific methodologies is the strand of hermeneutics. Derived from the Greek *hermeneuein* – to interpret or to explain – the term *hermeneutics* stands for a school, mainly associated with Hans-Georg Gadamer, which pursues the ideal of a human science of understanding. The emphasis is on interpretation in an interplay between a subject-matter and the interpreter's position. This emphasis introduces the subjective into scientific methodology. Hermeneutics denies both that a single "objective true interpretation" can transcend all individual viewpoints, and that humans are forever confined within their own ken [13]. This epistemology offers a necessary complement to a scientific stance, which exclusively hinges on "hard", quantitative methods in order supposedly to achieve absolute objectivity. The implication of hermeneutics for model validation is that it recognizes the pertinence of subjective judgment. In this connection, interpretive discourses play a crucial role in group model-building and validation. Such discourses lead beyond the subjective, entailing the creation of inter-subjective, shared realities. We will return to this factor in Sect. "Validation Process".

Critical Rationalism

Critical rationalism is a philosophical position founded by Karl R. Popper [19,20]. It grew out of positivism but rejected its verificationist stance. Critical rationalism posits that, in the social domain, theories can never be definitely proved, but can only reach greater or lesser levels of truth. Scientific proofs are confined to the realm of the formal sciences, namely logic and mathematics.

As Popper demonstrates, all theories are provisional. As a consequence, the main criterion for the assessment of a theory's truth status is *falsification* [19]. A theory holds as long as it is not refuted. Consequently, any theory can be upheld as long as it passes the test of falsification. In other words, the fertile approaches to science are not those of corroboration, but the falsificationist efforts to test if theories can be upheld. In the context of modeling this means that validation must undertake attempts to falsify a model, thereby testing its robustness.

Even Popper's theory of science is not unchallenged. For example, Kuhn has made the point that its principles are applicable only to normal science, which operates incrementally within a given paradigm, but not to anomalous science, which uncovers unsuspected phenomena in periods of scientific revolution [12]. This observation has an implication for model validation: Alternative and even multiple model designs should be assessed for their ability to account for fundamental change.

On the Meaning of Validity and Validation

One of the predominant convictions about science is the obsessive idea that proofs are the touchstone of the validity of both theories and models. We follow a different rationale, reverting to the philosophy of science as embodied in critical rationalism.

Popper's refutationist concept (as opposed to a verificationist concept) of theory-testing implies both an evolutionist perspective and an empiricist stance. The evolutionist perspective is primary because it welcomes the challenges posed to a theory, since these attempts at falsification lead to an evolutionary process: successful falsification efforts result in revisions and improvements of the theory. Correspondingly, empiricism is paramount in the social sciences, because the main source for the refutation of a theory is empirical evidence. However, falsification can also be grounded in logical arguments where empirical evidence cannot be obtained. In this sense, a structuralist approach as used in System Dynamics validation transcends the bounds of logical empiricism.

As a consequence of the evolutionist perspective, there is no such thing as absolute validity. Validity is always imperfect, but it can be improved over time. The empiricist aspect of theory-building implies that theories must be validated by means of empirical data. However, logical assay, estimation and judgment are complementary to this empiricist component (see below).

A validation process is about gradually building confidence in the model under study [2]. This is both analytical and synthetic. It is directed at the model as a whole as much as it is at the components of the model. The touchstone of validity is less whether the model is right or wrong: as Sterman states, "… all models are wrong." [28].

Some models, however, fulfill the purpose ascribed to them, i. e., they are useful. Models are inherently incomplete; they cannot claim to be true in an absolute sense, but only to be relatively true [4]. In this sense, *validation* is a *goal-oriented* activity and *validity* a *relative* concept.

Finally, the validation process often involves several people because the necessary knowledge is distributed. In these cases, the dialectics of propositions and refutations, as well as the interaction of different subjective viewpoints, and consensus-building, are integral. Validation processes, then, are semiformal, discursive social procedures with a holistic as opposed to a fragmentary orientation [ibidem].

On Objectivity

If subjective views and judgments are as prominent as alleged above, does objectivity play a role at all? Operational philosophy shows a way out of this dilemma: Rapoport defines objectivity as "invariance with respect to different observers." [21]. Popper has a similar stance in proposing that general statements must be formulated in a way that they can be criticized and, where applicable, falsified [20]. This concept of objectivity is a challenge to model validation: When defining concepts and functions, one must first of all strive for falsifiable statements. In principle, formal models meet this criterion: each variable and every function or relationship can be challenged. And they must be challenged, so that their robustness can be tested. The duty, then, is in finding the invariances that are inter-subjectively accepted as the best approximations to truth. Frequently this is best achieved in group model-building processes [30]. Finally, truth is something we search for but do not possess [20], i. e., even an accepted model cannot guarantee truth with final certainty.

Validation Methods

A considerable set of qualitative and quantitative tests has been developed for the enhancement of model validity. The state-of-the-art has been documented in seminal publications [2,4,7,8,14,17,28]. Our purpose here is to present and exemplify the different tests to encourage and help those who strive to develop high-quality System Dynamics models.

In the following, an overview of the types of tests developed for System Dynamics models is given, without any claim to completeness. Most of these tests have been documented extensively in [2,7,8,28]. The descriptions of the tests adhere closely to the specifications of these authors (mainly Forrester and Senge). In addition, we have developed a new category for tests that concentrate on the context in which the model is to be developed. High-quality models can be created only if the relevant context is taken into consideration. To facilitate orientation, we have attached an overview of all described tests in the Appendix.

In this section we describe three groups of tests: those related to model-related context, tests of model structure and tests of model behavior. Many of the tests described in the following can be utilized for explanatory analysis which aims at an understanding of the problematic behavior of the issue under study. Others are suitable for normative ends, in analyzes targeted on improvements of system performance with regard to a specified objective of the reference system. Also known as policy tests, or policy analyses, these "tests of policy implications differ from other tests in their explicit focus on comparing changes in a model and in the corresponding reality. Policy … tests attempt to verify that response of a real system to a policy change would correspond to the response predicted by the model" [8]. Policy testing can show the risk involved in adopting the model for policy making.

Tests About the Model-Related Context

These tests deal with aspects related to the situation in which the model is to be developed and embedded. They imply metalevel decisions which have to be taken in the first place, before engaging in model-building. Applied ex-post-facto, i. e., after modeling, they allow for assessing the utility of the modeling endeavor as such.

Issue Identification Test. The *raison d'être* of a System Dynamics model is its ability to adequately address an issue and to enhance stakeholders' understanding, an ability which may lead to policy insights and system improvements. The issue identification test examines whether or not the identified issue or problem is indeed meaningful. Has the "right" problem been identified? Does the problem statement address the origins of an issue or only superficial *symptoms*? Whenever complex issues are addressed by a model, different perspectives (e. g. professional, economic, political) must be integrated for accurate problem identification and modeling. This is not a "one-shot-only" test; it must be applied recurrently during the modeling procedure. By reflecting regularly on the correctness of the identified issue, the modeler can increase the likelihood of capturing the origins of suboptimal system behavior.

Adequacy of Methodology Test. Simulation models respond to the limitations of humans' mental ability to comprehend complex, dynamic feedback systems [27]. The adequacy of methodology test scrutinizes whether the

System Dynamics methodology is best-suited for dealing with the issue under study. One needs to clearly ascertain if that issue is characterized by dynamic complexity, feedback mechanisms, nonlinear interdependency of structural elements and delays between causes and effects. One needs to ask also if the issue under study could be better addressed by another methodology. For example, in a case where the question is to understand the difference in numerical outcomes between two configurations of a production system, it lets one determine whether discrete event simulation would fulfill this requirement more accurately than System Dynamics.

System Configuration Test. This test asks the fundamental question about whether the structural configuration chosen can be accepted. It challenges the assumption that the model represents the actual working of the system under study. The applicability of a different design would be suggested by its ability to capture new conditions, such as different system configurations, phenomena or rules of the game. Even revolutionary changes might be considered. Such an outlook may require a totally new model, or an alternative model designed from a different vantage point. This would at least feasibly approximate the need to take paradigmatic change into account.

System Improvement Test. The purpose of modeling is to understand a part of reality and to resolve an issue. The system improvement test can be performed only after the modeling project (an ex-post-facto test), once the insights derived from the model have already been implemented in the real system. This test reestablishes the connection between the abstract mathematical model and the real system. The system improvement test helps to evaluate whether or not model development was successful. In operational terms, any improvements of the real system under study must be compared with explicit objectives. In practice, the test might assess the impact of the modeling process or the model use either on the mental models of decision makers or on changes in organization structures. In principle, assessing the impact of a modeling endeavor is very difficult (one preliminary example is provided by Snabe and Grössler [25]).

Tests of Model Structure

Tests of model structure refer to the "nuts and bolts" of System Dynamics modeling, i. e., to the formal concepts and interrelationships which represent the real system. Model structure tests aim to increase confidence in the structure of the created theory about the behavior mode of interest. The model structure can be assessed by means of either direct or indirect inspection. Tests of model structure assess whether the logic of the model is attuned to the corresponding structure in the real world. They do not yet compare the model behavior with time series data from the real system.

Direct Structure Tests Direct structure tests assess whether or not the model structure conforms to relevant descriptive knowledge about the real system or class of systems under study. By means of direct comparison, they qualitatively assess any disparities between the original system structure and the model structure.

Structure Examination Test. Examination in this case means comparison in the sense just outlined. Qualitative or quantitative information about the real system structure can be obtained either empirically or theoretically. Empirically based tests include reviews of model assumptions about system elements and their interdependencies, e. g., reviews made by highly knowledgeable experts of the real system. Theory-based tests compare the model structure with theoretical knowledge from literature about the type of system being studied. Thereby, a preference for theoretical knowledge specific to the modeled situation over more abstract and general knowledge is usually the case.

To pass the structure examination test, a model must not contradict either the evidence or knowledge about the structure of the real system. This test ensures that the model contains only those structural elements and interconnections that are most likely extant in the real system. In this context, formal inspections of the model's equations, reviews of the syntax for the stock and flow diagram, and walkthroughs along the causal loop diagrams and their embodied causal explanations may be indicated. The experienced reader might recommend the use of statistical tests to identify and validate model structure. As Forrester and Senge [8] indicate, a long-standing discussion exists about the application of inferential statistical tests for structure examination. After a series of experiments, Forrester and Senge conclude "that conventional statistical tests of model structure are not sufficient grounds for rejecting the causal hypotheses in a system dynamics model." [8]. In the future, however, new statistical approaches might enrich the testing procedures.

Parameter Examination Test. A parameter is a quantity that characterizes a system and is held constant in a case under study, but may be varied in different cases (e. g., energy consumption per capita per day). The aim of parameter examination is to evaluate a model's parameters against evidence or knowledge about the real system. The test can utilize both empirical and theoretical information. Furthermore, the test can be conceptual or numerical. The conceptual parameter examination test is about construct validity; it identifies elements in the real system that correspond to the parameters of the model. Conceptual correspondence means that the parameters match elements of the real system's structure. Numerical parameter examination checks to see if the quantities of the conceptually confirmed parameters are estimated accurately. Techniques for the estimation of parameters are described in [9].

Direct Extreme Condition Test. Extreme conditions do not often occur in reality; they are exceptions. The validity of a model's equations under extreme conditions is evaluated by assessing the plausibility of the results generated by the model equations against the knowledge about what would happen under a similar condition in reality. Direct extreme condition testing is a mental process and does not involve computer simulation. Ideally, it is applied to each equation separately. It consists of assigning extreme values to the input variables of each equation. The values of the output variables are then interpreted in terms of what would happen in the real system under these extreme conditions. For example, if a population is zero, then neither births, deaths, nor consumption of resources can occur.

Boundary Adequacy Structure Test. Boundary adequacy is given if the model contains the relevant structural relationships that are necessary and sufficient to satisfy a model's purpose. Consequently, the boundary adequacy test inquires whether the chosen level of aggregation is appropriate and if the model includes all relevant aspects of structure. It should ensure that the model contains the concepts that are important for addressing the problem endogenously. For instance, if parameters are likely to change over time, they should be endogenized [8]. The pertinent validation question is: "Should this parameter be endogenized or not?" That question must be decided in view of the model's purpose.

The boundary adequacy test can be applied in three ways: as a structural test, as a behavioral test, and as a policy test. The names are correspondingly: boundary adequacy structure test, boundary adequacy behavior test, and boundary adequacy policy test.

As a test of model structure, the boundary adequacy test involves developing a convincing hypothesis relating the proposed model structure to the particular issue addressed by the model. The boundary adequacy behavior/policy test (explained in Subsect. "Indirect Structure Tests") continues this line of thinking.

Dimensional Consistency Test. This test checks the dimensional consistency of measurement units of the expressions on both sides of an equation. The test is performed only at the equation level. When all tests of the individual equations are passed, a large system of dimensionally consistent equations results. This test is passed only if consistency is achieved without the use of parameters that have no meaning in respect to the real world. The dimensional consistency test is a powerful test to establish the internal validity of a model.

Indirect Structure Tests Indirect structure tests assess the validity of the model structure indirectly by examining model-generated outcome behaviors. These tests require computer simulation. The comparative activities in these tests are based on logical plausibility considerations which in turn are based on the mental models of the analyst. Comparisons of model generated data and time series about the real system are not yet involved. The tests can be applied to different degrees of model completeness, i. e., to the smallest "atomic" model components, to sub-models, as well as to the entire model.

Indirect Extreme Condition Test. For this test, the modeler assigns extreme values to selected model parameters and compares the generated model behavior to the observed or expected behavior of the real system under the same extreme conditions. This test is the logical continuation of the direct extreme condition test, i. e., many of the extreme conditions mentally developed in the previous stage can now be deployed to evaluate the simulated behavioral consequences. This test can be used for the explanatory analysis phase of modeling, but also for the normative phase of policy development. In the first instance, indirect extreme conditions are used to develop a structure that can reproduce the system behavior of interest and guard against developments impossible in reality. In the latter instance, the introduction of policies aims to improve the system's performance. The indirect extreme policy test introduces extreme policies to the model and compares the simulated consequences to what would be the most likely outcome of the real system if the same extreme policies would have been implemented.

Behavior Sensitivity Test. Sensitivity analysis assesses changes of model outcome behavior given a systematic variation of input parameters. This test reveals those parameters to which the model behavior is highly sensitive, and asks if the real system would exhibit a similar sensitivity to changes in the corresponding parameters. "The behavior sensitivity test examines whether or not plausible shifts in model parameters can cause a model to fail behavior tests previously passed. To the extent that such alternative parameter values are not found, confidence in the model is enhanced." [8]. A model can be numerically sensitive, i. e., the numerical values of variables change significantly, but the behavioral patterns are conserved. It can also exhibit behavioral sensitivity, i. e., the modes of model behavior change remarkably based on systematic parameter variations (Barlas [3] defines several distinct patterns of model behavior).

As the test for indirect extreme conditions, the behavior sensitivity test can also be deployed to assess policy sensitivity. It can reveal the degree of robustness of model behavior and hence indicate to what degree model-based policy recommendations might be influenced by uncertainty in parameter values. If the same policies would be recommended regardless of parameter changes over a plausible range, risk in using the model would be lower than if two plausible sets of parameters lead to distinct policy recommendations.

Integration Error Test. Integration error is the deviation between the analytical solution of differential equations and the numerical solution of difference equations. This test ascertains whether the model behavior is sensitive to changes in either the applied integration method or the chosen integration interval (often referred to as simulation time step). Euler's method is the simplest numerical technique for solving ordinary differential and difference equations. For models that require more precise integration processes, the more elaborated Runge–Kutta integration methods can produce more accurate results, but they require more computational resources.

Boundary Adequacy Behavior Test/Boundary Adequacy Policy Test. The logic for testing boundary adequacy has already been developed under the aspect of direct structure testing in the preceding section. The indirect structure version of this test asks whether model behavior would change significantly if the boundary were extended or reduced; i. e., the test involves conceptualizing additional structure or canceling unnecessary structure with regard to the purpose of the study. As one example of expanding the model boundary, this version of the test allows one to detail the treatment of model assumptions considered as unrealistically simple but still important for the model's purpose. On the other hand, simplifying the model is also a way to reduce the model boundary. The loop-knockout analysis is a useful method to implement this two-sided test. Knockout analysis checks behavior changes induced by the connection and disconnection of a portion of the model structure, and helps the modeler to evaluate the usefulness of those changes with respect to the model's purpose.

The other version of this test is the boundary adequacy policy test. It examines whether policy recommendations would change significantly if the boundary were extended (or restricted): That is, what would happen if the boundary assumptions were relaxed (or confined)?

Loop Dominance Test. Loop dominance analysis studies the internal mechanisms of a dynamic model and their temporal, relative contribution to the outcome behavior of the model. The relative contribution of a mechanism is a complex quantitative statement that explains the fraction of the analyzed behavior mode caused by the mechanism considered in ► System Dynamics, Analytical Methods for Structural Dominance Analysis in. The analysis reveals the relative strengths of the feedback loops in the model. The loop dominance test compares these results with the modeler's or client's assumption about which are the dominant feedback loops in the real system. Since the results are analytical statements, interpretation and comparison with the real system requires profound knowledge about the system under study.

Loop dominance analysis reveals insights about a model on a different level of analysis than the other validation tests discussed so far: It works not on the level of individual concepts or behaviors of variables but on the level of causal structure, and compares the temporal significance of the different structures to each other. The use of this test for model validation is a novelty. If the relative loop dominances of the model map the relative loop dominances of the real system, confidence in the model is enhanced. If the relative loop dominances of the real system are not known, it is still possible to evaluate whether or not the loop dominance logic in the model is reasonable.

Tests of Model Behavior

Tests of model behavior are empirical and compare simulation outcomes with data from the real system under study. On that basis, inferences about the adequacy of the model can be made. The empirical data can either be historical or refer to reasonable expectations about possible future developments.

Behavior Reproduction Tests The family of behavior reproduction tests examines how well model-generated behavior matches the observed historical behavior of the real system. As a principle, models should be tested against data not only from periods of stability but also from unstable phases. Policies should not be designed or tested on the premise of normality, but rather should be validated with a view toward robustness and adaptiveness.

Symptom Generation Test. This test indicates whether or not a model produces the symptom of difficulty that motivated the construction of the model. To pass the symptom generation test is a prerequisite for considering policy changes, because "unless one can show how internal policies and structures cause the symptoms, one is in a poor position to alter those causes" [8].

Summary statistics, which measure and enable the interpretation of quantitative deviations, provide the means to operationalize the symptom generation test.

One known example is Theil inequality statistics, which measures the mean square-error (MSE) between the model-generated behavior and the historical time series data. It breaks down the deviation into three sources of error: Bias (U_m), unequal variation (U_s), and unequal covariation (U_c) [26].

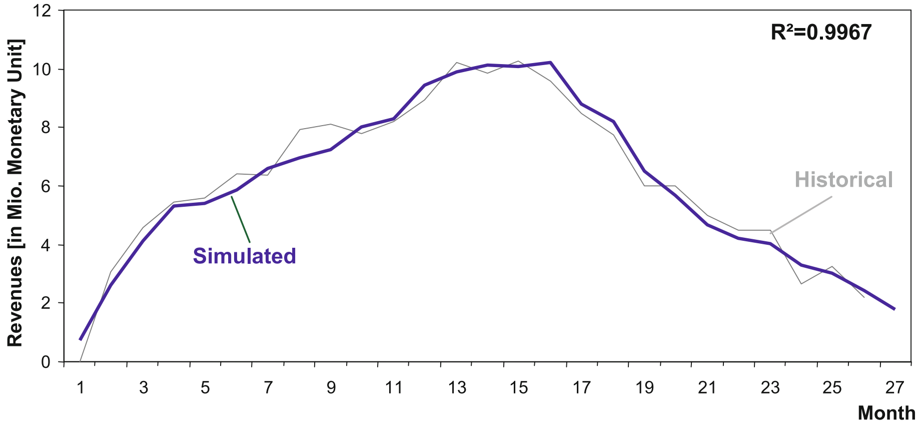
An example taken from Schwaninger and Groesser [22] illustrates the interpretation of the error sources.

This example from an industrial firm concerns the design of a model that replicates the observed, historical product life-cycle pattern with high accuracy (Fig. 2). "Product Revenue" is the main variable of interest and specifies the symptom (growth phase followed by rapid decay). The mean square-error for revenues is 0.35. The individual components of the inequality statistics are: $U_m = 0.01, U_s = 0.01, U_c = 0.98$. The break down of the statistics shows that the major part of the error is in the U_c component, while the other two sourees of error are small. This signifies that the point-by-point values of the simulated and historical data do not match, even though the model captures the dominant trend and the average values in the historical data. Such a situation indicates that a major part of the error is probably unsystematic, and therefore the model should not be rejected for failing to match the noise component of the data. The residuals of the historic and simulated time series show no significant trend. This strengthens the assessment that the model comprises a structure that captures the fundamental dynamics of the issue under consideration.

Frequency Generation and Phase Relationship Tests. These tests focus on the frequencies of time series and phase relationships between variables. An example is the pattern of investment cycles in an industry. These tests are superior to point-by-point comparisons between model-generated and observed behavior (cf. [7]).

Frequency refers to periodicities of fluctuation in a time series. Phase relationship is the relationship between the time series of at least two variables. In principle, three phase relations are possible: Preceding, simultaneous, and successive. The frequency generation test evaluates whether or not the periodicity of a variable is in accordance with the real system. The phase relationship test assesses the phase shifts of at least two variables by comparing their trajectories.

If the phase shift between the selected simulation variables contradicts the phase shift between the same variables as observed or expected in the real system, a structural flaw in the model might be diagnosed. The test can



System Dynamics Modeling: Validation for Quality Assurance, Figure 2
An example comparison of historical and simulated time series for product revenues. The explained variance is close to 100% ($R^2 = 0.9967$)

uncover failures in the model, but offers only little guidance as to where the erroneous part of the model might be. The autocorrelation function test is one way to operationalize the frequency generation test [1]. The function test consists in comparing the autocorrelation functions of the observed and the model-generated behavior outputs, and can detect if significant errors between them exist.

Modified Behavior Test. Modified behavior can arise from a modified model structure or changes in parameter values. This test concerns changes in the model structure. It can be performed if data about the behavior of a structurally modified version of the real system are available. "The model passes this test if it can generate similar modified behavior, when simulated with structural modifications that reflect the structure of the "modified" real system" [2]. The applicability of this test is rather limited since it requires specific data about the modified real system which must be similar in kind to the original real system. Only under this condition can additional insights into the suitability of the original model structure be obtained. If the modified real system deviates strongly from the original real system, the test does not result in any additional insights, because no stringent conclusions about the validity of the original system can be derived from a model that is dissimilar in its structure.

Multiple Modes Test. A mode is a pattern of observed behavior. The multiple mode test considers whether a model is able to generate more than one mode of observed behavior, for instance, if a model about the production sector of an economy generates distinct patterns of fluctuations for the short-term (production, employment, inventories, and prices) and for the long term (investment, capital stock) [15]. " A model able to generate two distinct periodicities of fluctuation observed in a real system provides the possibility for studying possible interaction of the modes and how policies differentially affect each mode" [8].

Behavior Characteristic Test. Characteristics of a behavior are features of historical data that are clearly distinguishable, e. g., the peculiar shape of an oscillating time series, sharp peaks, long troughs, or such unusual events as an oil crisis. Since System Dynamics modeling is not about point prediction, the behavior characteristic test evaluates whether or not the model can generate the circumstances and behavior leading to the event. The creation of the exact time of the behavior is not part of the test.

Behavior Anticipation Tests System Dynamics models do not strive to forecast future states of system variables. Nevertheless, given that the fundamental system structure is not subject to rapid and fundamental change, dynamic models might provide insights about the possible range of future behaviors. Hence, behavior anticipation tests are similar to behavior reproduction tests but possess a higher level of uncertainty.

Pattern Anticipation Test. This test examines whether a model generates patterns of future behavior which are assumed to be qualitatively correct. The limits of anticipation reside in the fact that that the structure of the system may change over time. The pattern anticipation test entails evaluation of periods, phase relationships, shape, or other characteristics of behavior anticipated by the model. One possibility for implementing this test is to split the historical time series into two data sets and introduce an artificial present time at the end of the first data series. The first set is then used for model development and calibration. The second data series is employed to perform the behavior anticipation test, i. e., to evaluate whether the model is able to anticipate possible future behavior.

This test can also be used for policy considerations, in which case it is called "Changed Behavior Anticipation Test". It determines whether the model correctly anticipates how the behavior of the real system will change if a governing policy is altered.

Event Anticipation Test. In respect to System Dynamics, the anticipation of events does not imply knowing the exact time at which the events occur; it rather means understanding the dynamic nature of events and being able to identify the antecedents leading to them. For instance, the event anticipation test is passed if a model has the ability to anticipate a steep peak in food prices based on the development of the conditioning factors.

Behavior Anomaly Test In constructing and analyzing a System Dynamics model, one strives to make it behave like the real system under study. However, the analyst may detect anomalous features of the model's behavior which conflict with the behavior of the real system. Once the behavioral anomaly is traced to components of the model structure responsible for the anomaly, one often finds flaws in model assumptions. The test for recognizing behavioral anomalies is sporadically applied throughout the modeling process.

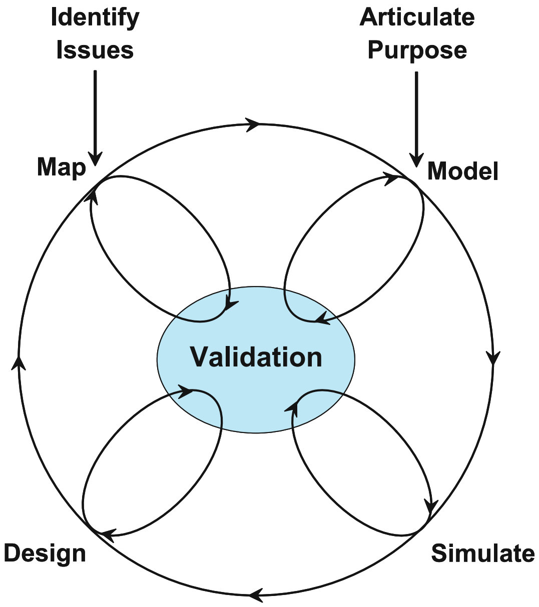
Family Member Test A System Dynamics model often represents a family of social systems. Whenever possible, a model should be a general representation of the class of that system to which the particular case belongs. One should ask if the model can generate the behavior in other instances of the same class. "The family-member test permits a repeat of the other tests of the model in the context of different special cases that fall within the general theory covered by the model. The general theory is embodied in the structure of the model. The special cases are embodied in the parameters. To perform this test, one uses the particular member of the general family for picking parameter values. Then one examines the newly parametrized model in terms of the various model tests to see if the model has withstood transplantation to the special case" [8]. The model should be calibrated so as to be applicable to the widest range of related systems. For the family member test, only the parameter values of the model are subject to alterations; changes in the model structure are part of the modified behavior test, as discussed in the preceding section.

Surprise Behavior Test A surprising model behavior is a behavior that is not expected by the analysts. When such an unexpected behavior appears, the model analysts must first understand the causes of the unexpected behavior within the model. They then compare the behavior and its causes with those of the real system. In many cases, the surprising behavior turns out to be due to a formulation flaw in the model. However, if this procedure leads to the identification of behavior previously unrecognized in the real system, the confidence in the model's usefulness is strongly enhanced. Such a situation may signify a model-based identification of a counter-intuitive behavior in a social system.

Turing Test The Turing test is a qualitative test which uses the intuitive knowledge of system experts to evaluate model behavior. Experts are presented with a shuffled collection of real and simulated output behavior patterns. They are asked if they can distinguish between these two types of patterns. If they are unable to discern which pattern belongs to the real system and which to the simulation output, the Turing test is passed. Similar to the phase relationship test, the Turing test is powerful in its ability to indicate structural flaws, but offers only little guidance for locating them in the model.

Validation Process

The validation process pervades all phases of model-building and reaches even beyond, into the phases of model implementation and use. The diagram in Fig. 3 visualizes the function of validation in the process of model-building.



System Dynamics Modeling: Validation for Quality Assurance, Figure 3
Validation in the context of the System Dynamics modeling procedure

For the purposes of this contribution, validation is placed at the center of the scheme. From there it is dispersed through all steps of the modeling process, Map (high-level model creation), Model (build the formal model), Simulate (explore scenarios, etc.) and Design (articulation of policies). We have limited the differentiation of these steps in order to highlight the structure of the process – a recursive structure drawn as a nested loop line. After the initial identification of issues and the articulation of model purpose, the simplified diagram denotes the four phases, of mapping to modeling to simulation and design. The small loops symbolize micro-processes in which, for example, a model is submitted to validation, e. g., a direct structure test, which may lead to its modification (two small arrows). The larger loops illustrate more comprehensive processes. For example, an indirect structure test of the model is carried out, in which the behavior is tested by means of simulation. Or a policy test by simulation leads to implications for design (large loop), and the design is validated in detail thereafter (small loop).

Now, we should note that the process scheme reminds us of a further aspect which is quite fundamental. If the results of the model's operation, e. g., a "prediction", diverge from the results of a test, then either the model is wrong or the test is inadequate (see p. 168 in [24]). This meta-perspective lets us keep an eye on the adequacy of the tests:

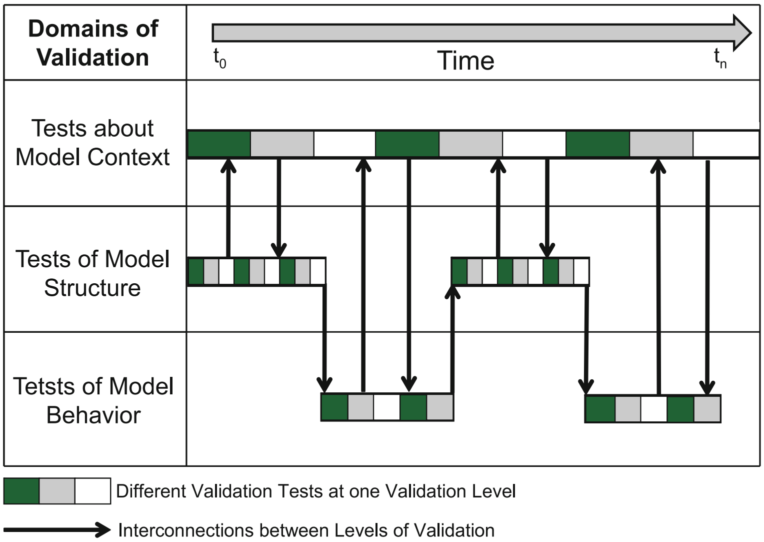
is the logic of the test flawless? Are the data sources in order? (see adequacy of methodology test in Sect. "Validation Methods").

Model-building is a process of knowledge-creation, and model validation is an integral part of it. As the model is validated using the methods described in the former chapter, insights emerge, and a better understanding of the system under study keeps growing. But model-building is also a construction of a reality in the minds of observers [31,32] concerned with an issue. In this procedure, validation is supposed to be a "guarantor" for the realism of the model, a control function for preventing gross aberrations in individual and collective perceptions. Validation should encompass precautions against cognitive limitations and modeler blindness. The set of tests presented above is a system of heuristic devices for enhancing such provisions. A question not yet answered is how these tests should be ordered along the timeline. We have fleshed out three structural principles, which are illustrated in Fig. 4:

1. *Validation is a parallel process:* Validation in all three domains – context, structure and behavior – is carried out in a synchronized fashion, as shown in Fig. 4. Context validation is continuous, while the other two components show alternations.
2. *Parts of the validation process have a sequential structure:* This refers to the alternations between the components of structure and behavior validation. In principle, they occur alternately, with structural validation taking the lead and behavior validation following. After that, one might revert to structural validation again and so forth.
3. *Validation processes are polyrhythmic:* The length and accentuation of validation activities vary among the three levels. This fact is symbolized by the frequency of the vertical lines in the blocks of the chronogram.

A further important factor affecting the validation process is the degree of resolution: micro, meso or macro (as visualized in Fig. 1). The focus of validation is primarily on micro-objects, the smallest building blocks of a model, for example, a stock or a subsystem containing a stock with its flows. One could call them metaphorically *atoms* or *molecules*. Each building-block should be validated individually, before it is integrated into the overall model structure. The reason is that at this atomic level disfunctionalities or errors of thinking are discovered immediately, while at higher levels of resolution the identification of structural flaws is more difficult and cumbersome. The same holds for the relation between modules (meso) and the whole model (macro). Before adding a module, it should be validated in itself. This way, errors at the level of the whole system can be minimized and, it is very important to add, counterintuitive behavior of the model can be understood with more ease.

Until now we have examined what occurs in a validation process and how the process is structured. Finally, we raise the issue of who the actors are and why. In this context, we will concentrate on group processes in model validation.


Domains of Validation
Time
t_0
t_n
Tests about Model Context
Tests of Model Structure
Tetsts of Model Behavior
Different Validation Tests at one Validation Level
Interconnections between Levels of Validation

System Dynamics Modeling: Validation for Quality Assurance, Figure 4
The interplay of validation activities

Different observers associate diverse contents with a system, and they might even conceive the system distinctly, as far as its boundaries, goals and structures are concerned. They might also succumb to erroneous inferences and therefore adhere to defective propositions. Consequently, error-correcting devices are needed. A powerful mechanism for this purpose is the practice of model-building and validation in groups. We have already referred to that concept in respect to several of the methods discussed in Sect. "Validation Methods", and now we will briefly expand on it.

Group Model-building (GMB) is a methodology to facilitate team learning with the help of System Dynamics [30]. The methodology consists of a set of methods and instruments as well as heuristic principles. These are meant to facilitate the elicitation of knowledge, the negotiation of meanings, the creation of a shared understanding of a problem in a team, as well as the joint construction and validation of models. The process of GMB is essentially a dialog in which different interpretations of the real system under study are exposed, transformed, aligned and translated into the concepts and relationships which make up the model system. This is mainly a matter of structural validation, of qualitative mapping and the elaboration of the formal model.

Given its transdisciplinary approach, GMB enables an integration of different perspectives into one shared image of the system-in-focus. GMB is an important provision for attaining higher model quality: it can broaden the available knowledge base, inhibit errors and show itself to be a cohesive force in the quest for consensual model validation. The opportunity for validation inheres in the broad knowledge base normally available in a modeling group. Much of this knowledge can be leveraged for validation purposes. Most validation tests are carried out in coordination with model-building activities. Often the tests become a task to be accomplished between workshops. However, the members of the model-building group can, in principle, be made available for knowledge input into and monitoring of validation activities.

A functioning GMB process requires a number of necessary elements [18]: commitment of key players (e. g., attendance of workshops), impartial facilitation, on-the-spot modeling at conversational pace, with continuous display of the developing model as well as an interactive and iterative group process.

Let us not forget that there are many situations in which one single person is in charge of building and validating a model. In these cases the modeler must constantly challenge his or her own position. Normally, it is preferred that one should also call for external judgment in reviews, walkthroughs and the like. The same holds for knowledge supply. One-person modelers can find a lot of material in the media, libraries, the internet, etc., but it is also usually beneficial to find experienced persons from whom to elicit relevant knowledge, or even persons who join the modeling and validation venture.

Synopsis and Outlook

Models should be relevant for coping with the complexity of the real world. At the same time, the methods by which they are constructed must be rigorous; otherwise the quality of the model suffers. Rigor and relevance are not entirely dichotomous, but given resource constraints they are in competition to a certain extent. Lack of rigor in building a model is often worse than limitations to the model's relevance. One may say, *cum grano salis*: incomplete validation entails complete irrelevance. Modelers must find a way to ensure both rigor and relevance, as both are necessary conditions for achieving the model purpose. Neither alone is sufficient, but one may assume that, taken together, rigor and relevance are sufficient conditions. The relative importance of these two dimensions of model building may vary over time as a function of the model quality achieved. At the beginning, relevance might be more important, while at high levels of model accomplishment rigor might become prevalent.

Investing in high model quality is indeed both worthwhile and imperative. It is impressive to register the fact that model validation has achieved higher levels of rigor not only in the academic field but also in the world of affairs: According to Coyle and Exelby, the need for orientating decisions about "real-world" affairs has also fueled strong efforts among commercial modelers and consultants for ensuring model validity [5].

We have discussed two essential aspects of model validation, the epistemological foundations and methodological procedures for ensuring model validity. The main conclusion we have reached on epistemology is that crude positivism has been superseded by newer philosophical orientations that provide guidance for an adequate concept of validation in System Dynamics. Validation has been defined as a rich and well-defined process by which the confidence in a model is gradually enhanced. Validity, then, is always a matter of degree, never an absolute property.

Well-defined here is not meant in the sense of a rigid algorithm, but as the rigorous application of a battery of validation methods which we have described in some detail. We have included a number of new validation tests by which modelers' understanding of the relevant context can be scrutinized. These additional tests are rightly supposed to prevent wrong methodological choices. They should also trigger innovative approaches to the issues under study and foster the ability to think in terms of contingencies. Finally, they should liberate modelers from tunnel vision and open avenues to creativity. The imperative here is to cultivate a "sense of the possible" (Robert Musil's *Möglichkeitssinn*) and a skepticism against the supposedly impossible (see also [29]).

Simulation based on formal dynamic models is likely to become ever more important for both private and public organizations. It will continue to support managers at all levels in decision-making and policy design. The more that models are relied upon, the greater the importance of their high quality. Therefore, model validation is one of the big issues lying ahead in System Dynamics modeling.

Appendix: Overview of the Tests Described in This Chapter

1. **Tests of the Model-Related Context**
    - 1.1 Issue Identification Test
    - 1.2 Adequacy of Methodology Test
    - 1.3 System Configuration Test
    - 1.4 System Improvement Test
2. **Tests of Model Structure**
    - 2.1 Direct Structure Tests
        - 2.1.1 Structure Examination Test
        - 2.1.2 Parameter Examination Test
        - 2.1.3 Direct Extreme Condition Test
        - 2.1.4 Boundary Adequacy Structure Test
        - 2.1.5 Dimensional Consistency Test
    - 2.2 Indirect Structure Tests
        - 2.2.1 Indirect Extreme Condition Test
        - 2.2.2 Behavior Sensitivity Test
        - 2.2.3 Integration Error Test
        - 2.2.4 Boundary Adequacy Behavior Test/Boundary Adequacy Policy Test
        - 2.2.5 Loop Dominance Test
3. **Tests of Model Behavior**
    - 3.1 Behavior Reproduction Tests
        - 3.1.1 Symptom Generation Test
        - 3.1.2 Frequency Generation and Phase Relationship Test
        - 3.1.3 Modified Behavior Test
        - 3.1.4 Multiple Modes Test
        - 3.1.5 Behavior Characteristic Test
    - 3.2 Behavior Anticipation Tests
        - 3.2.1 Pattern Anticipation Test
        - 3.2.2 Event Anticipation Test
    - 3.3 Behavior Anomaly Test
    - 3.4 Family Member Test
    - 3.5 Surprise Behavior Test
    - 3.6 Turing Test

Bibliography

Primary Literature

1. Barlas Y (1990) An autocorrelation function test for output validation. Simulation 55(1):7–16
2. Barlas Y (1996) Formal aspects of model validity and validation in system dynamics. Syst Dyn Rev 12(3):183–210
3. Barlas Y (2006) Model validity and testing in System Dynamics: Two specific tools. Paper presented at the 24th International Conference of the System Dynamics Society, Nijmegen
4. Barlas Y, Carpenter S (1990) Philosophical roots of model validity – two paradigms. Syst Dyn Rev 6(2):148–166
5. Coyle G, Exelby D (2000) The validation of commercial system dynamics models. Syst Dyn Rev 16(1):27–41
6. Feyerabend P (1993) Against method, 3rd edn. Verso, London
7. Forrester JW (1961) Industrial dynamics. MIT Press, Cambridge
8. Forrester JW, Senge PM (1980) Test for building confidence in System Dynamics models. In: Legasto AA Jr, Forrester JW, Lyneis JM (eds) System Dynamics. North-Holland Publishing Company, Amsterdam, pp 209–228
9. Graham AK (1980) Parameter estimation in system dynamics modeling. TIMS Studies in the Management Sciences 14:125–142
10. Heracleous L (2006) Discourse, interpretation, organization. Cambridge University Press, Cambridge
11. James W (1987) Writings 1902–1910. Library of America, New York
12. Kuhn T (1996) The structure of scientific revolutions, 3rd edn. University of Chicago Press, Chicago
13. Lacey AR (1996) A dictionary of philosophy, 3rd revised edn. Barnes and Noble, New York
14. Lane DC (1995) The folding star: A comparative reframing and extension of validity concepts in System Dynamics. In: Simada T, Saeed K (eds) Proceedings of 1995 international System Dynamics conference, 30 July–4 Aug, vol I. System Dynamics Society, Lincoln, pp 111–130
15. Mass NJ (1975) Economic cycles: An analysis of underlying causes. Productivity Press, Cambridge
16. Mattheij RMM, Rienstra SW, Boonkkamp JH MtT (2005) Partial differential equations: Modeling, analysis, computation. Society for Industrial and Applied Mathematics (SIAM), Eindhoven
17. Petersen DW, Eberlein RL (1994) Understanding models with vensim. In: Morecroft JDW, Sterman JD (eds) Modeling for learning organiziations. Productivity Press, Portland, pp 339–358
18. Phillips LD (2007) Decision conferencing. In: Edwards W, Miles RF, von Winterfeldt D (eds) Advances in decision analysis. From foundations to applications. Cambridge University Press, Cambridge, pp 375–399
19. Popper KR (1959) The logic of scientific discovery. Basic Books, New York (latest edition: 2002, Routledge, London)
20. Popper KR (1972) Objective knowledge: An evolutionary approach. Clarendon Press, Oxford
21. Rapoport A (1954) Operational philosophy. Integrating knowledge and action. Harper, New York

22. Schwaninger M, Groesser SN (2008) Model-based theory-building with system dynamics. Syst Res Behav Sci 25:1–19
23. Seiffert H, Radnitzky G (1994) Handlexikon der Wissenschaftstheorie, 2nd edn. DTV Wissenschaft, Munich
24. Smith VL (2008) Rationality in economics: Constructivist and ecological forms. Cambridge University Press, Cambridge
25. Snabe B, Grössler A (2006) System dynamics modelling for strategy implementation – case study and issues. Syst Res Behav Sci 23(4):467–481
26. Sterman JD (1984) Appropriate summary statistics for evaluating the historical fit of system dynamics models. Dynamica 10(2):51–66
27. Sterman JD (1989) Misperceptions of Feedback in Dynamic Decision Making. Organ Behav Human Decis Process 43(3):301–335
28. Sterman JD (2000) Business dynamics. Systems thinking and modeling for a complex world. Irwin/McGraw-Hill, Boston
29. Taleb NN (2007) The black swan. The impact of the highly improbable. Random House, New York
30. Vennix JAM (1996) Group model building: Facilitating team learning using System Dynamics. Wiley, Chichester
31. von Foerster H (1984) Observing systems, 2nd edn. Intersystems Publications, Seaside
32. von Glasersfeld E (1991) Abschied von der Objektivität. In: Watzlawick P, Krieg P (eds) Das Auge des Betrachters. Piper, Munich, pp 17–30

Books and Reviews

Finlay PN (1997) Validity of decision support systems: Towards a validation methodology. Syst Res Behav Sci 14(3):169–182
Forrester JW (1961) Industrial dynamics. MIT Press, Cambridge
Law AM (2007) Simulation modeling and analysis, 4th edn. McGraw-Hill, New York
Legasto AA Jr, Forrester JW, Lyneis JM (eds) (1980) System Dynamics. North-Holland, Amsterdam
Morecroft J (2007) Strategic modelling and business dynamics: A feedback systems approach. Wiley, Chichester
Sargent RG (2004) Validation and verification of simulation models. In: Ingalls RG, Rossetti MD, Smith JS, Peters BA (eds) Proceedings of the 2004 winter simulation conference. ACM-Association for Computing Machinery, Washington DC, pp 17–28
Sterman JD (2001) Business dynamics. Systems thinking and modeling for a complex world. Irwin/McGraw-Hill, Boston
Warren K (2008) Strategic management dynamics. Wiley, Chichester

System Dynamics Models of Environment, Energy and Climate Change

Andrew Ford
School of Earth and Environmental Sciences,
Washington State University, Pullman, Washington, USA

Article Outline

Glossary

CO_2 Carbon dioxide is the predominant greenhouse gas. Anthropogenic CO_2 emissions are created largely by the combustion of fossil fuels.

CGCM Coupled general circulation model, a climate model which combines the atmospheric and oceanic systems.

GCM General circulation model, a term commonly used to describe climate models maintained at large research centers.

GHG GHG is a greenhouse gas such as CO_2 and methane. These gases contribute to global warming by capturing some of the outgoing infrared radiation before it leaves the atmosphere.

GT Gigaton, a common measure of carbon storage in the global carbon cycle. A GT is a billion metric tons.

IPCC The Intergovernmental Panel on Climate Change was formed in 1988 by the World Meteorological Organization and the United Nations Environmental Program. It reports research on climate change. Their assessments are closely watched because of the requirement for unanimous approval by all participating delegates.

Definition of the Subject

System dynamics is a methodology for studying and managing complex systems which change over time. The method uses computer modeling to focus our attention on the information feedback loops that give rise to the dynamic behavior. Computer simulation is particularly useful when it helps us understand the impact of time delays and nonlinearities in the system. A variety of modeling methods can aid the manager of complex systems. Coyle (p. 2 in [3]) puts the system dynamics approach in perspective when he describes it as that “branch of control theory which deals with socio-economic systems, and that branch of management science which deals with problems of controllability." The emphasis on controllability can be traced to the early work of Jay Forrester [9] and his background in control engineering [10]. Coyle highlighted controllability again in the following, highly pragmatic definition:

> *System dynamics is a method of analyzing problems in which time is an important factor, and which involve the study of how a system can be defended against, or made to benefit from, the shocks which fall upon it from the out-side world.*

The emphasis on controllability is important as it directs our attention to understanding and managing the system, not to the goal of forecasting the future state of the system. Making point predictions is the objective of some modeling methods, but system dynamics models are used to improve our understanding of the general patterns of dynamic behavior. System dynamics has been widely used in business, public policy and energy and environmental policy making. This article describes applications to energy and environmental systems.

Introduction

System dynamics has been used extensively in the study of environmental and energy systems. This article describes some of these applications, paying particular attention to the problem of global climate change. The applications were selected to illustrate the power of the method in promoting an interdisciplinary understanding of complex problems.

The applications to environmental and energy systems are similar to applications to other systems described in this encyclopedia. They usually begin with the recognition of a dynamic pattern that represents a problem. System dynamics is based on the premise that we can improve our understanding of the dynamic behavior by the construction and testing of computer simulation models. The models are especially helpful when they illuminate the key feedbacks that give rise to the problematic behavior.

System dynamics is explained in the core article in this volume, in the early texts by Forrester [9], Coyle [3] and Richardson [18] and in more recent texts on strategy by Warren [22] and by Morecroft [17]. The most comprehensive explanation is provided in the text on business dynamics by Sterman [19]. Applications to environmental systems are explained in the text by Ford [7]. The most widely read application to the environment is undoubtedly *The Limits to Growth* [16]. Collections of environmental applications appear in special issues of the *System Dynamics Review* [11,20].

The models are normally implemented with visual software such as Stella (http://www.iseesystems.com), Vensim (http://www.vensim.com/) or Powersim (http://www.powersim.com/). These programs use stock and flow icons to help one see where the accumulations of the system take place. They also help one to see the information feedback in the simulated system. The programs use numerical methods to show the dynamic behavior of the simulated system. The examples selected for this article make use of the Stella and Vensim software.

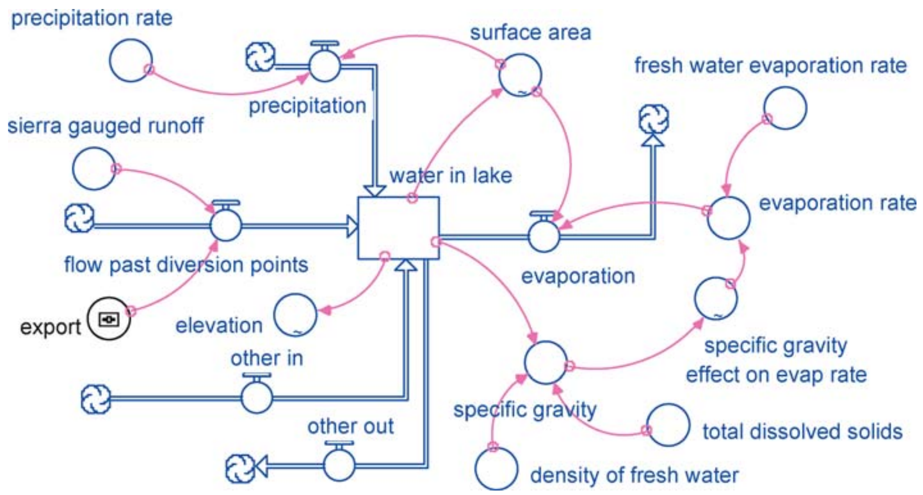
This article begins with textbook examples of environmental resources in the western US. The management of water levels at Mono Lake in Northern California is the first example. It shows a hydrological model to simulate the decline in lake levels due to water exported out of the basin. The second example involves the declining salmon population in the Tucannon River in Eastern Washington. These examples demonstrate the clarity of the approach, and they illustrate the potential for interdisciplinary modeling.

The article then turns to the topic of climate change and global warming. The focus is on the global carbon cycle and the growing concentration of carbon dioxide (CO_2) in the atmosphere. A wide variety of models have been used to improve our understanding of the climate system and the importance of anthropogenic CO_2 emissions. Examples of system dynamics models are presented to show how they can improve our understanding and provide a platform for interdisciplinary analysis.

System dynamics has also been widely applied in the study of energy problems, especially problems in the electric power industry. The final section describes two applications to electric power. The first involved the financial problems of regulated electric utilities in the US during the 1970s. It demonstrates the usefulness of the method in promoting an interdisciplinary understanding of the utilities' financial problems. The second study dealt with the CO_2 emissions in the large electricity system in the Western USA and Canada. It demonstrated how the power industry could lead the way in reducing CO_2 emissions in the decades following the implementation of a market in carbon allowances.

The Model of Mono Lake

Mono Lake is an ancient inland sea on the east side of the Sierra Nevada Mountains in California. Microscopic algae thrive in its saline waters, and the algae support huge populations of brine flies and brine shrimp which can,



System Dynamics Models of Environment, Energy and Climate Change, Figure 1
Stella diagram of the model of Mono Lake

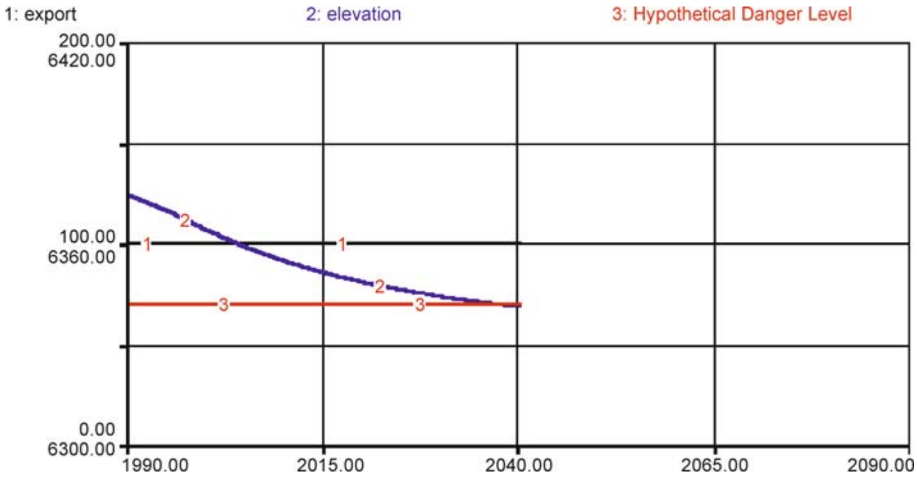
under the right conditions, provide a virtually limitless food supply for migratory and nesting birds. Starting in 1941, stream flows toward Mono Lake were diverted into the aqueduct for export to Los Angeles. The large export deprived the lake of the historical flows, and the volume shrunk over the next four decades. By 1980, the lake's volume was cut approximately in half, and its salinity nearly doubled. Higher salinity levels posed risks to the ecosystem, and environmental scientists feared for the future of the lake ecosystem. Various groups filed suit in the 1970s to limit exports, and the California Supreme Court ruled in 1983 that public trust doctrine mandated a reconsideration of the management of the waters of the Mono Basin. That reconsideration led to a long-term plan to limit exports until the lake's elevation would return to safer levels.

Figure 1 shows a system dynamics model to simulate water flows and storage in the Mono Basin. The goal was to understand the pattern of decline over four decades and to study the responsiveness of the lake to a change in export policy. The model is implemented with the Stella software, and Fig. 1 shows how the model appears when using the software. A single stock variable is used to represent the storage in the basin. The main flow into the lake is the flow from gauged streams that bring runoff from the Sierra to the lake. The aqueduct system diverts a portion of this flow south to Los Angeles, and the flow allowed past the diversion points is the main flow into the lake. The main outflow is the evaporation. It depends on the surface area of the lake and the evaporation rate. The surface area depends in a nonlinear way on the volume of water in the lake. Figure 1 shows that this model follows the standard, system dynamics practice of using familiar names to convey the meaning of the variables in the model. (These particular names match the terms used by water managers and hydrological models of the basin.)

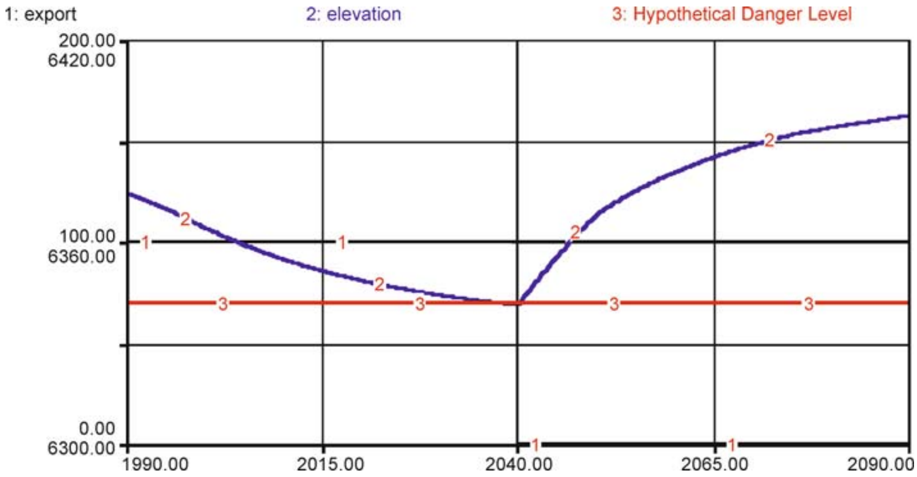
Figure 2 shows the simulated decline in the lake if exports were allowed to continue at high levels for 50 years. The lake would decline from 6374 to around 6342 feet above sea level, a value which is designated as a hypothetical danger level for this simulation. The long, gradual decline is a match of projections by the other hydrological models used in the management plan for the basin. The lake will continue to fall until the area has been reduced sufficiently to create an evaporation which will lead to a balance of the flows in and out of the basin.

Figure 3 shows the simulated responsiveness of the lake to a change in export. The export is cut to zero midway through the simulation, and the elevation increases rapidly in the ensuing decade. The simulation reveals an immediate and rapid response, indicating that there is little downward momentum associated with the hydrology of the basin. This responsiveness is highly relevant to the management plan. When the lake falls to a dangerous level, the export could be reduced, and the lake would climb to higher elevations within a few years after the change in policy. This rapid response supports the "wait and see" argument by those who advocated waiting for full signs of a dangerous salinity before changing export policy.[1] But there is far more than hydrology at work in this system. The waters of Mono Lake support a com-

[1] "Wait and see" may be supported by an analysis of the hydrology of the basin, but it does not necessarily make sense when considering the long delays in the political and managerial process to change water export.



System Dynamics Models of Environment, Energy and Climate Change, Figure 2
Simulated decline in Mono Lake elevation if historical export were allowed to continue until the year 2040

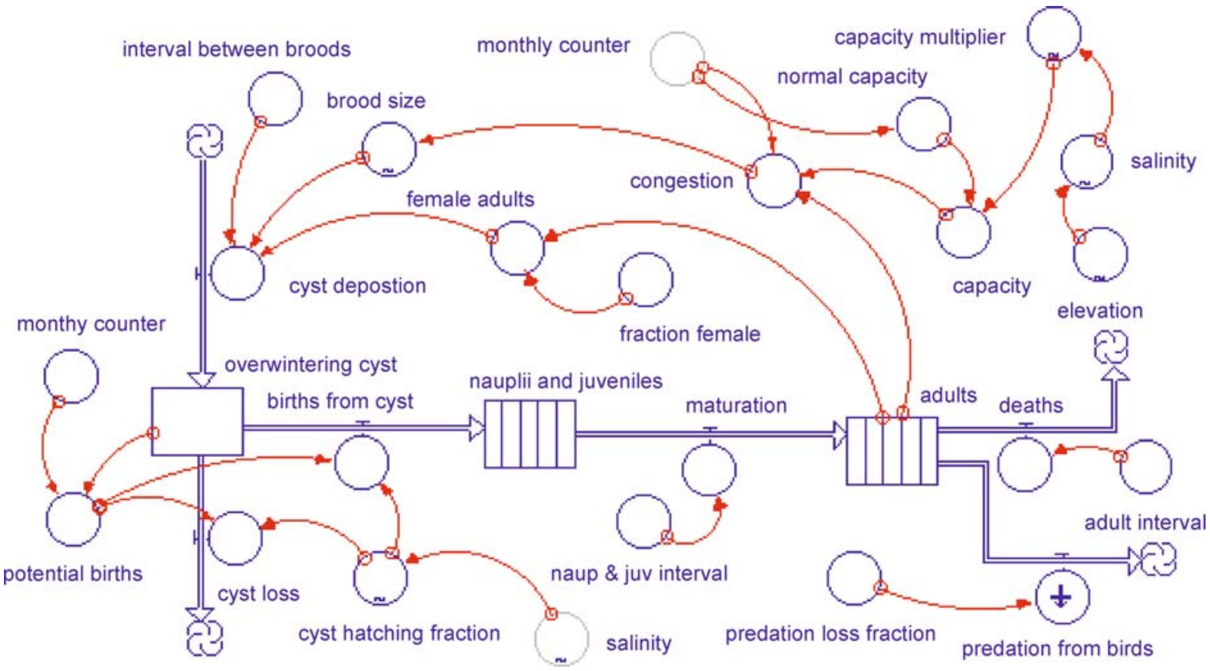


System Dynamics Models of Environment, Energy and Climate Change, Figure 3
Simulated recovery of Mono Lake elevation if export is set to zero for the second half of the simulation

plex ecosystem which may or may not recover as quickly as the lake elevation. To explore the larger system requires an interdisciplinary model, one that looks at both hydrology and population biology.

Figure 4 shows a model of the population of brine shrimp that live in Mono Lake. The life cycle begins when the adult females deposit cysts in the summer. A stock is assigned to the over wintering cysts. The nauplii and juvenile phases are combined into a second stock, and the maturation leads to a new population of adults in the following summer. The model operates in months and is simulated over a long time interval to show the population response to long-term changes in elevation and in salinity. The model shows the population's response to changes in lake elevation, so one can learn about the delays in the population's response to the changes in lake elevation. Since the shrimp life cycle is 12 months, one would expect the population to rebound rapidly after the increase in elevation and the reduction in salinity. The model confirms that the shrimp population would increase rapidly in the years following the elimination of water export from the basin.

The Mono Lake models are textbook models [7]. They demonstrate the clarity that the system dynamics approach brings to the modeling of environmental systems. The stock and flow icons help one see the structure of the system, and the long variable names help one appreciate the individual relationships. The simulation results help one understand the downward momentum in the system. In this particular case, there is no significant down-



System Dynamics Models of Environment, Energy and Climate Change, Figure 4
Stella model of the brine shrimp population of Mono Lake

ward momentum associated with either the hydrological dynamics or the population dynamics.

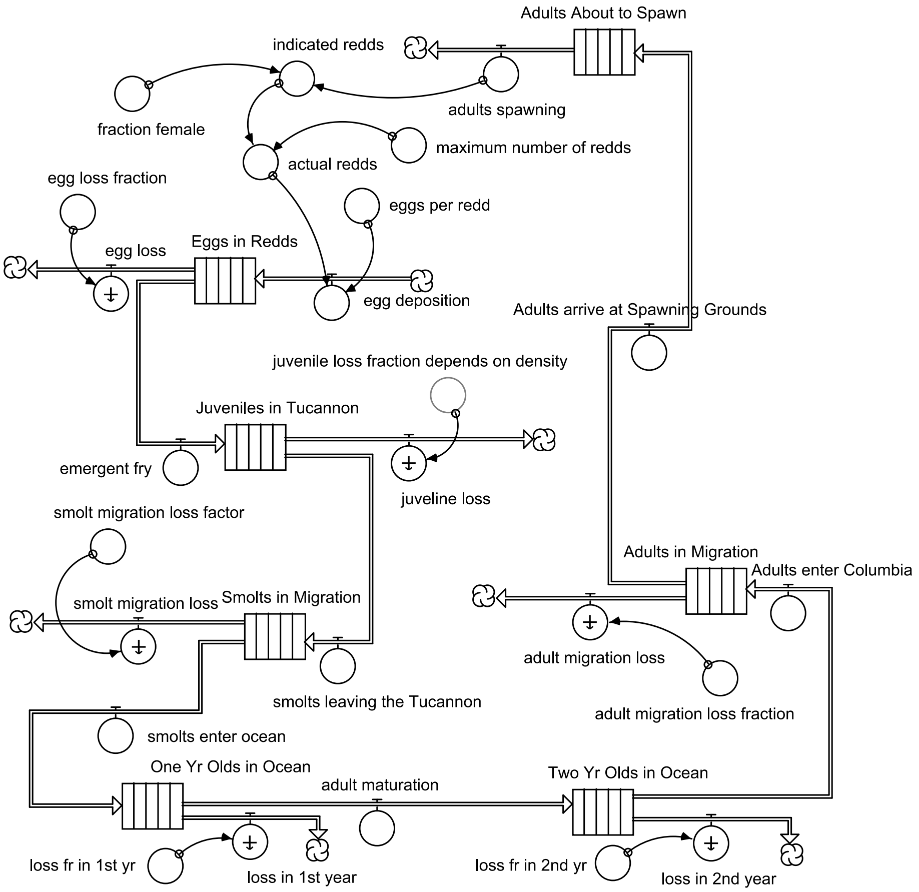
The model in Fig. 1 allows for a system dynamics portrayal of the type of calculations commonly performed by hydrologists. Compared to the previous methods in hydrology, system dynamics adds clarity and ease of experimentation. The population model in Fig. 4 is a system dynamics version of the type of modeling commonly performed by population biologists. System dynamics adds clarity and ease of experimentation in this discipline as well.

The main theme of this article is that system dynamics offers the opportunity for interdisciplinary modeling and exploration. The Mono Lake case illustrates this opportunity with the combination of the hydrological and biological models that allows one to simulate management policies that control export based on the size of the brine shrimp population. The new model is no longer strictly hydrology nor strictly population biology; it is an interdisciplinary combination of both. And by using stock and flow symbols that are easily recognized by experts from many fields of study, the system dynamics enables quick transfer of knowledge. The ability to combine perspectives from different disciplines is one of the most useful aspects of the system dynamics approach to environmental and energy systems. This point is illustrated further with each of the remaining examples in the article.

The Model of the Salmon in the Tucannon River

The next example involves the decline in salmon populations in the Snake and Columbia River system of the Pacific Northwest. By the end of the 1990s, the salmon had disappeared from 40% of their historical breeding ranges despite a public and private investment of more than $1 billion. The annual salmon and steelhead runs had dwindled to less than a quarter of the runs from one hundred years ago. Figure 5 shows a system dynamics model one of the salmon runs, the population of Spring Chinook that spawn in the Tucannon River. The river rises in the Blue Mountains of Oregon and flows 50 miles toward the Snake River in Eastern Washington. It is estimated that the river originally supported runs of 20 thousand adults. But the number of returning adults has declined substantially due to many changes in the past sixty years. These changes include agricultural development in the Tucannon watershed, hydro-electric development on the Snake and Columbia, and harvesting in the ocean.

Each of the stocks in Fig. 5 correspond to a different phase in the salmon life cycle (see Table 1), with a total lifecycle of 48 months. The parameters represent predevelopment conditions, the conditions prior to agricultural development in the Tucannon watershed and hydro-electric development on the Snake and Columbia. Each of these parameters is fixed regardless of the size of the salmon



System Dynamics Models of Environment, Energy and Climate Change, Figure 5
Stella diagram of the model of the salmon life cycle

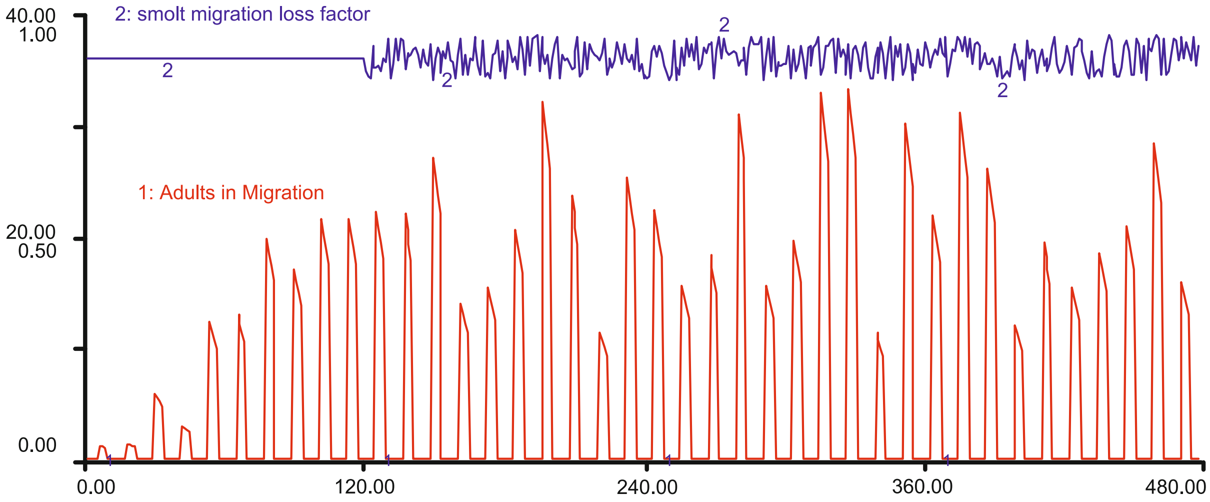
populations. One of the most important variables is the "juvenile loss fraction depends on density." It can be as low as 50% when there are only a few emergent fry each spring. With higher densities, however, juvenile survival becomes more difficult due to crowding in the cool and safe portions of the river.

Figure 6 shows the model results over a 480 month period with the population parameters in Table 1. The simulation begins with a small number to see if the population will grow to the 20 thousand adults that were thought to have returned to the river in earlier times. The time graph shows a rapid rise to around 20 thousand adults within the first 120 months of the simulation. The remainder of the simulation tests the population response to variability in environmental conditions, as represented by random variations in the smolt migration loss fraction. (This loss tends to be high in years with low runoff and low in years with high runoff.) Figure 6 confirms that the model simulates the major swings in returning adults due to environmental variability. The runs can vary from a low of ten thousand to a high of thirty thousand.

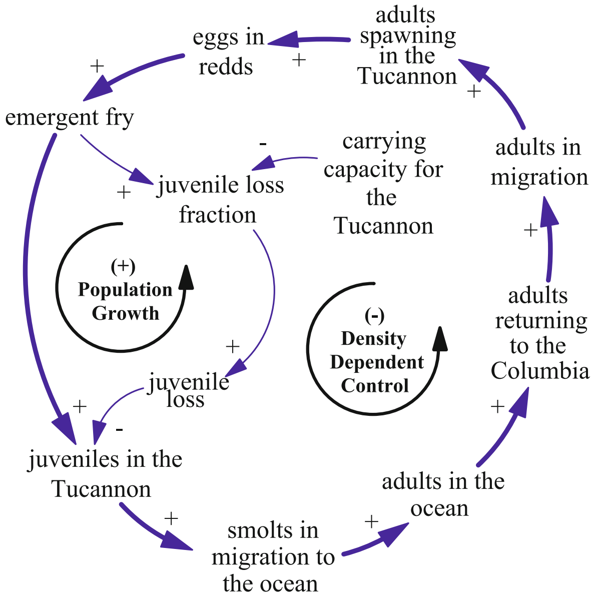
System dynamics models are especially useful when they help us to understand the key feedbacks in the system. Positive feedback loops are essential to our under-

System Dynamics Models of Environment, Energy and Climate Change, Table 1
Inputs to simulate the salmon population under pre-development conditions

Months in each phase		Population parameters	
Adults ready to spawn	1	fraction female	50%
eggs in redds	6	eggs per redd	3,900
juveniles in Tucannon	12	egg loss fraction	50%
smolts in migration	1	smolt migration loss factor	90%
one yr olds in ocean	12	loss fr for first yr	35%
two yr olds in ocean	12	loss fr for second yr	10%
adults in migration	4	adult migration loss fraction	25%



System Dynamics Models of Environment, Energy and Climate Change, Figure 6
Test of the salmon model with random variations in the smolt migration losses


eggs in redds
adults spawning in the Tucannon
emergent fry
carrying capacity for the Tucannon
adults in migration
juvenile loss fraction
(+) Population Growth
(-) Density Dependent Control
adults returning to the Columbia
juvenile loss
juveniles in the Tucannon
adults in the ocean
smolts in migration to the ocean
+
-

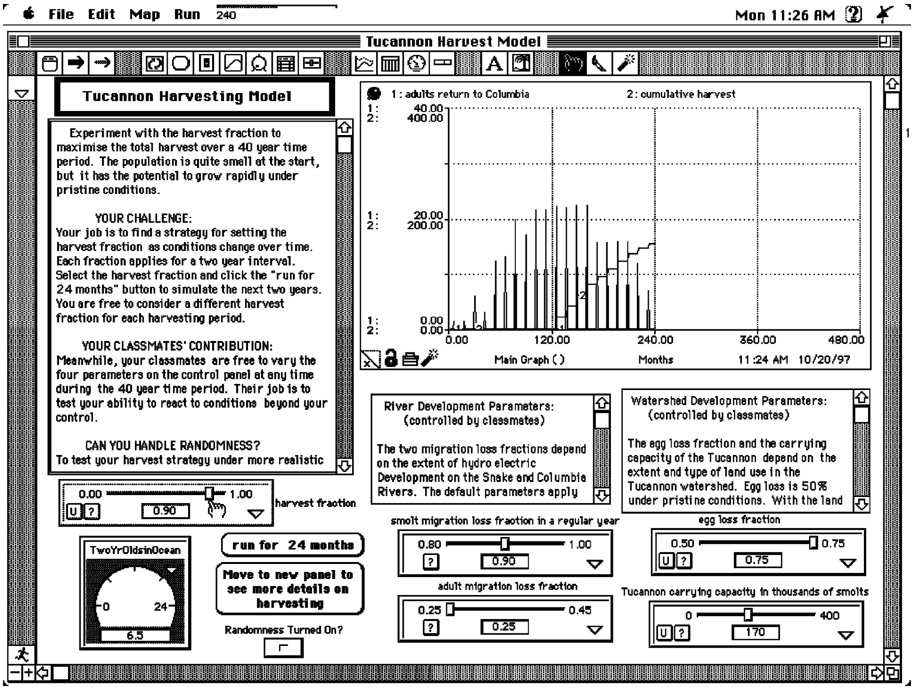
System Dynamics Models of Environment, Energy and Climate Change, Figure 7
Key feedback loops in the salmon model

standing of rapid, exponential growth; negative feedbacks are essential to our understanding of the controllability of the system. Causal loop diagrams are often used to depict the feedback loops at work in the simulated system. Figure 7 shows an example by emphasizing the most important feedback loops in the salmon model.

Most readers will immediately recognize the importance of the outer loop which is highlighted by bold arrows in the diagram. Starting near the top, imagine that there are more spawning adults and more eggs in redds. We would then expect to see more emergent fry, more juveniles, more smolts in migration, more salmon in the ocean, more adults entering the Columbia, and a subsequent increase in the number of spawning adults. This is the positive feedback loop that gives the salmon population the opportunity to grow rapidly under favorable conditions.

An equally important feedback works its way around the inner loop in the diagram. If we begin at the top with more spawners, we would expect to see more eggs, more fry and a greater juvenile loss fraction as the fry compete for space in the river. With a higher loss fraction, we expect to see fewer juveniles survive to be smolts, fewer smolts in migration, and fewer adults in the ocean. This means we would see fewer returning adults and less egg deposition. This “density dependent feedback” becomes increasingly strong with larger populations, and it turns out to be crucial to the eventual size of the population. Simulating density dependent feedback is also essential to our understanding of the recovery potential of the salmon population. Suppose, for example, that the salmon experience high losses during the adult migration, This will mean that fewer adults reach the spawning grounds. There will be less egg deposition and fewer emergent fry in the following spring. The new cohort of juveniles will then experience more favorable conditions, and a larger fraction will survive the juvenile stage and migrate to the ocean. The density dependent feedback is crucial to the population’s ability to withstand shocks from external conditions.[2]

[2] The shocks could take the form of changes in ocean mortalities, changes in harvesting and changes in the migration mortalities. These



Tucannon Harvesting Model

Experiment with the harvest fraction to maximise the total harvest over a 40 year time period. The population is quite small at the start, but it has the potential to grow rapidly under pristine conditions.

YOUR CHALLENGE:
Your job is to find a strategy for setting the harvest fraction as conditions change over time. Each fraction applies for a two year interval. Select the harvest fraction and click the "run for 24 months" button to simulate the next two years. You are free to consider a different harvest fraction for each harvesting period.

YOUR CLASSMATES' CONTRIBUTION:
Meanwhile, your classmates are free to vary the four parameters on the control panel at any time during the 40 year time period. Their job is to test your ability to react to conditions beyond your control.

CAN YOU HANDLE RANDOMNESS?
To test your harvest strategy under more realistic

River Development Parameters:
(controlled by classmates)

The two migration loss fractions depend on the extent of hydro electric Development on the Snake and Columbia Rivers. The default parameters apply

Watershed Development Parameters:
(controlled by classmates)

The egg loss fraction and the carrying capacity of the Tucannon depend on the extent and type of land use in the Tucannon watershed. Egg loss is 50% under pristine conditions. With the land

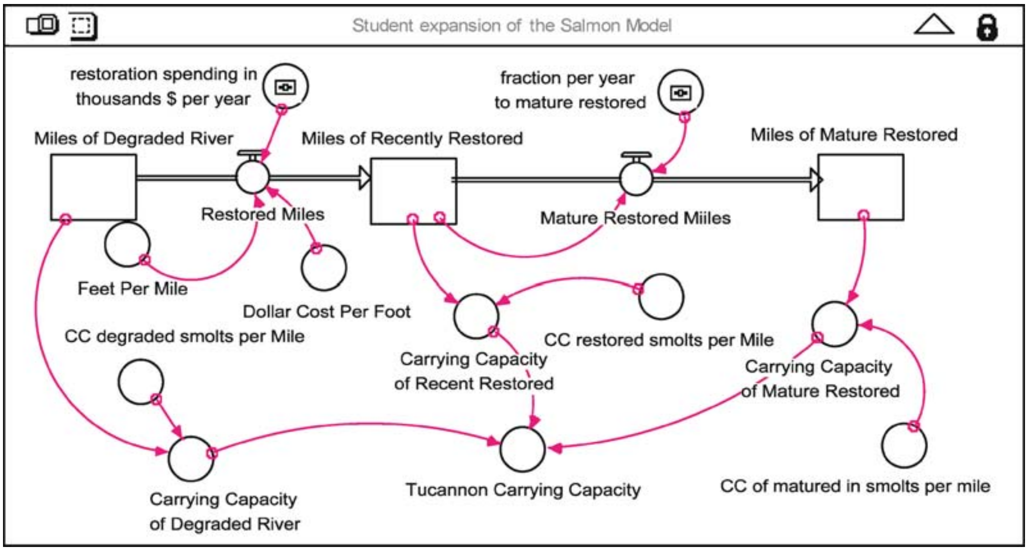
System Dynamics Models of Environment, Energy and Climate Change, Figure 8
Salmon harvesting model to encourage student experimentation

Figure 8 shows a version of the model to encourage student experimentation with harvesting policies. The information fields instruct the students to work in groups of three with one student playing the role of "the harvest manager". The harvest manager's goal is to achieve a large, sustainable harvest through control of the harvest fraction. The other students are given control of the parameters that describe conditions on the Snake and Columbia and in the Tucannon watershed. These students are encouraged to make major and unpredictable changes to test the instincts of the harvest manager.

Models designed for highly interactive simulations of this kind are sometimes called "management flight simulators" because they serve the same function as actual flight simulators. With a pilot simulator, the trainee takes the controls of an electro-mechanical model and tests his instincts for managing the simulated airplane under difficult conditions. The Tucannon harvesting model provides a similar opportunity for environmental students. They can learn the challenge of managing open access fisheries that are vulnerable to over harvesting and the tragedy of the commons [12]. In this particular exercise, students learn that they can achieve a sustainable harvest under a wide variety of difficult and unpredictable conditions. The key to sustainability is harvest manager's freedom to change the harvest fraction in response to recent trends in number of returning adults. This is an important finding for fishery management because it reveals that the population dynamics are not the main obstacle to sustainability. Rather, unsustainable harvesting is more likely to occur when the managers find it difficult to change the harvest fraction in response to recent trends. This is the fundamental challenge of an open-access fishery.

The salmon model is a system dynamics version of the type of modeling commonly performed by population biologists. System dynamics adds clarity and ease of experimentation compared to these models. It also provides a launching point for model expansions that can go beyond population biology. Figure 9 shows an example. This is a student expansion to change the carrying capacity from a user input to a variable that responds to the user's river restoration strategy. The student was trained in geomorphology and was an expert on restoring degraded rivers in the west. The Tucannon began the simulation with 25 miles of river in degraded condition and the remaining 25 miles in a mature, fully restored river with a much higher carrying capacity. The new model per-

shocks are external to the boundary of this model, so one is reminded of Coyle's definition of system dynamics. That is, the model helps us understand how the salmon population could withstand the shocks which fall upon it from the out-side world.


Student expansion of the Salmon Model
restoration spending in thousands $ per year
fraction per year to mature restored
Miles of Degraded River
Miles of Recently Restored
Miles of Mature Restored
Restored Miles
Mature Restored Miiles
Feet Per Mile
Dollar Cost Per Foot
CC degraded smolts per Mile
CC restored smolts per Mile
Carrying Capacity of Recent Restored
Carrying Capacity of Mature Restored
Carrying Capacity of Degraded River
Tucannon Carrying Capacity
CC of matured in smolts per mile

System Dynamics Models of Environment, Energy and Climate Change, Figure 9
Student addition to simulate river restoration

mits one to experiment with the timing of river restoration spending and to learn the impact on the management of the salmon fishery.

The student's model provides another example of interdisciplinary modeling that aids our understanding of environmental systems. In this particular case, the modeling of river restoration is normally the domain of the geomorphologist. The model of the salmon population is the domain of the population biologist. Their work is often conducted separately, and their models are seldom connected. This is unfortunate as the experts working in their separate domains miss out on the insights that arise when two perspectives are combined within a single model. In the student's case, surprising insights emerged when the combined model was used to study the economic value of the harvesting that could be sustained in the decades following the restoration of the river. To the student's surprise, the new harvesting could "pay back" the entire cost of the river restoration in less than a decade.

Models of Climate Change

Scientists use a variety of models to keep track of the greenhouse gasses and their impact on the climate. Some of the models combine simulations of the atmosphere, soils, biomass and ocean response to anthropogenic emissions. The more developed models include CO_2, methane, nitrous oxides and other greenhouse gas (GHG) emissions and their changing concentrations in the atmosphere. Claussen [2] classifies climate models as simple, intermediate and comprehensive. The simple models are sometimes called "box models" since they represent the storage in the system by highly aggregated stocks. The parameters are usually selected to match the results from more complicated models. The simple models can be simulated faster on the computer, and the results are easier to interpret. This makes them valuable for sensitivity studies and in scenario analysis [13].

The comprehensive models are maintained by large research centers, such as the Hadley Center in the UK. The term "comprehensive" refers to the goal of capturing all the important processes and simulating them in a highly detailed manner. The models are sometimes called GCMs (general circulation models). They can be used to describe circulation in the atmosphere or the ocean. Some simulate both the ocean and atmospheric circulation in a simultaneous, interacting fashion. They are said to be coupled general circulation models (CGCMs) and are considered to be the "most comprehensive" of the models available [2]. They are particularly useful when a high spatial resolution is required. However, a disadvantage of the CGCMs is that only a limited number of multi-decadal experiments can be performed even when using the most powerful computers.

Intermediate models help scientists bridge the gap between the simple and the comprehensive models. Claussen [2] describes eleven models of intermediate complexity. These models aim to "preserve the geographic integrity of the Earth system" while still providing the opportunity for multiple simulations to "explore the parameter space with some completeness. Thus, they are more suitable for assessing uncertainty". Figure 10 characterizes the different categories of models based on their relative emphasis on:

- number of processes (right axis)
- detailed treatment of the each process (left axis), and the
- extent of integration among the different processes (top axis).

Regardless of the methodology, climate modeling teams must make some judgments on where to concentrate their attention. No model can achieve maximum performance along all three dimensions. (Figure 10 uses the dashed lines to draw our attention to the impossible task of doing every thing within a single model.)

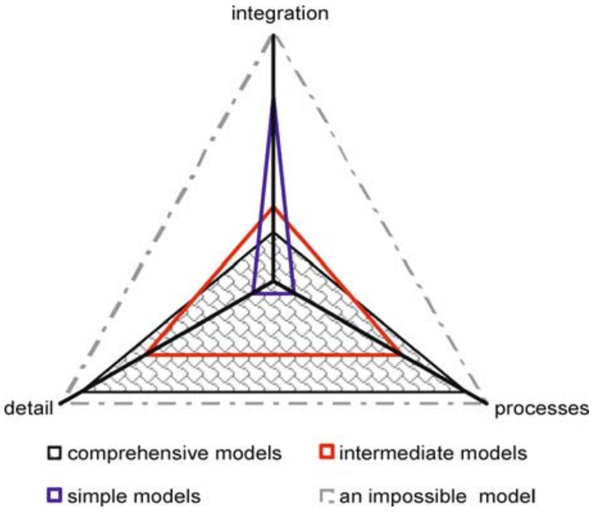
The comprehensive models strive to simulate as many processes as possible with a high degree of detail. This approach provides greater realism, but the models often fail to simulate the key feedback loops the link that atmospheric system with the terrestrial and oceanic systems. (An example is the feedback between CO_2 emissions, temperatures and the decomposition of soil carbon. If higher temperatures lead to accelerated decomposition, the soils could change from a net sink to a net source of carbon [15].) The simple models sacrifice detail and the number of processes in order to focus on the feedback effects between the processes. Using Claussen's terminology, one would say that such models aim for a high degree of "integration". However, the increased integration is achieved by limiting the number of processes and the degree of detail in representing each of the processes.

System dynamics has been used in a few applications to climate change. These applications fit in the category of simple models whose goal is to provide a highly integrated representation of the system. Two examples are described here; both deal with the complexities of the global carbon cycle.

System Dynamics Models of the Carbon Cycle

Figures 11 and 12 depict the global carbon cycle. Figure 11 shows the carbon flows in a visual manner. Figure 12 uses the Vensim stock and flow icons to summarize carbon storage and flux in the current system. The storage is measured in GT, gigatons of carbon, (where carbon is the C in CO_2). The flows are in GT/year of carbon with values rounded off for clarity.

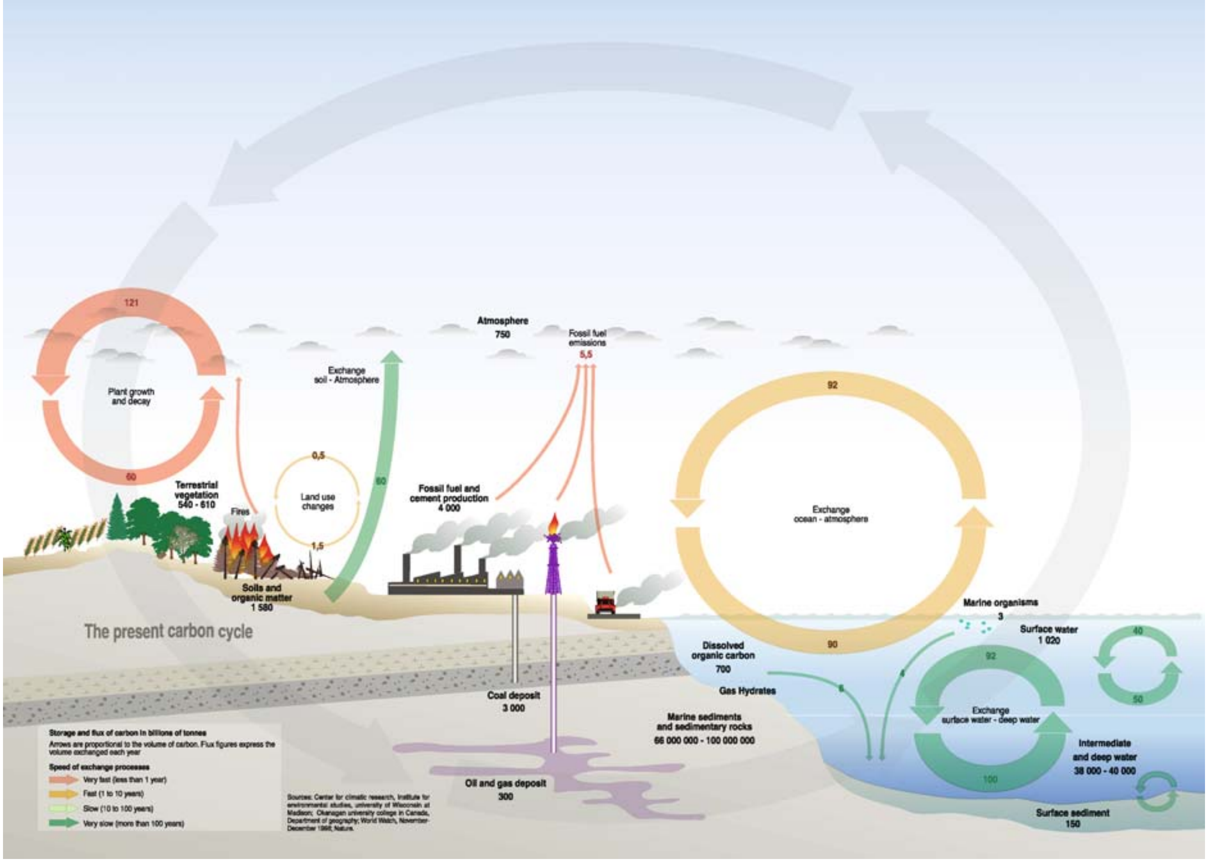
The left side of Fig. 12 shows the flows to the terrestrial system. The primary production removes 121 GT/yr from the atmosphere. This outflow exceeds the return flows by 1 GT/year. This imbalance suggests that around 1 GT of carbon is added to the stocks of biomass and soil each year. So the carbon stored in the terrestrial system would grow over time (perhaps due to extensive reforestation of previously cleared land.) The right side of Fig. 12 shows the flows from the atmosphere to the ocean. The CO_2 dissolved in the ocean each year exceeds the annual release back to the atmosphere by 2 GT. The total, net-flow out of the atmosphere is 3 GT/year which means that natural processes are acting to negate approximately half of the current anthropogenic load.



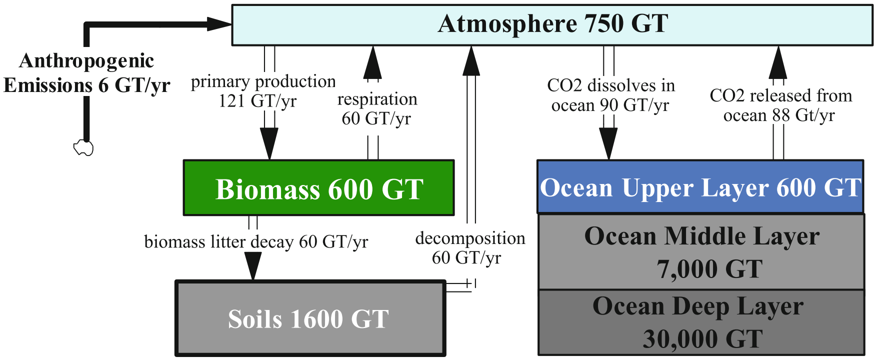
System Dynamics Models of Environment, Energy and Climate Change, Figure 10
Classification of climate models

As the use of fossil fuels grows over time, the anthropogenic load will increase. But scientists do not think that natural processes can continue to negate 50% of an ever increasing anthropogenic load. On the terrestrial side of the system, there are limits on the net flow associated with reforestation of previously cleared land. And there are limits to the carbon sequestration in plants and soils due to nitrogen constraints. On the ocean side of the system, the current absorption of 2 GT/year is already sufficiently high to disrupt the chemistry of the ocean's upper layer. Higher CO_2 can reduce the concentration of carbonate, the ocean's main buffering agent, thus affecting the ocean's ability to absorb CO_2 over long time periods.

Almost of the intermediate and comprehensive climate models may be used to estimate CO_2 accumulation in the atmosphere in the future. For this article, it is useful to draw on the mean estimate published in *Climatic Change* by Webster [23]. He used the climate model developed at the Massachusetts Institute of Technology, one of the eleven models of "intermediate complexity" in the


121
Plant growth and decay
60
Terrestrial vegetation 540 - 610
Fires
Land use changes
0,5
1,5
Exchange soil - Atmosphere
60
Atmosphere 750
Fossil fuel emissions 5,5
Fossil fuel and cement production 4 000
Soils and organic matter 1 580
The present carbon cycle
Coal deposit 3 000
Oil and gas deposit 300
92
Exchange ocean - atmosphere
90
Marine organisms 3
Surface water 1 020
Dissolved organic carbon 700
Gas Hydrates
6
4
92
Exchange surface water - deep water
100
40
50
Marine sediments and sedimentary rocks 66 000 000 - 100 000 000
Intermediate and deep water 38 000 - 40 000
Surface sediment 150
Storage and flux of carbon in billions of tonnes
Arrows are proportional to the volume of carbon. Flux figures express the volume exchanged each year
Speed of exchange processes
Very fast (less than 1 year)
Fast (1 to 10 years)
Slow (10 to 100 years)
Very slow (more than 100 years)
Sources: Center for climatic research, Institute for environmental studies, university of Wisconsin at Madison; Okanagan university college in Canada, Department of geography; World Watch, November-December 1998; Nature.

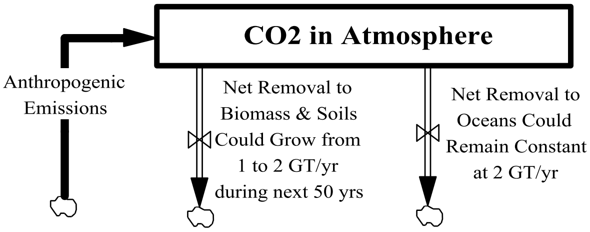
System Dynamics Models of Environment, Energy and Climate Change, Figure 11
The global carbon cycle. (Source: United Nations Environmental Program (UNEP) http://www.unep.org/)


Atmosphere 750 GT
Anthropogenic Emissions 6 GT/yr
primary production 121 GT/yr
respiration 60 GT/yr
CO2 dissolves in ocean 90 GT/yr
CO2 released from ocean 88 Gt/yr
Biomass 600 GT
Ocean Upper Layer 600 GT
biomass litter decay 60 GT/yr
decomposition 60 GT/yr
Ocean Middle Layer 7,000 GT
Soils 1600 GT
Ocean Deep Layer 30,000 GT

System Dynamics Models of Environment, Energy and Climate Change, Figure 12
Diagram of the stocks and flows in the carbon cycle

review by Claussen [2]. The model began the simulation in the year 2000 with an atmospheric CO_2 concentration of 350 parts per million (ppm). (This concentration corresponds to around 750 GT of carbon in the atmosphere.) The mean projection assumed that anthropogenic emissions would grow to around 19 GT/year by 2100. The mean projection of atmospheric CO_2 was around 700 ppm by 2100. The amount of CO_2 in the atmosphere would be twice as high at the end of the century.

Figure 13 shows the simplest possible model to explain the doubling of atmospheric CO_2. The stock accumulates the effect of three flows, each of which is specified by the


CO2 in Atmosphere
Anthropogenic Emissions
Net Removal to Biomass & Soils Could Grow from 1 to 2 GT/yr during next 50 yrs
Net Removal to Oceans Could Remain Constant at 2 GT/yr

System Dynamics Models of Environment, Energy and Climate Change, Figure 13
Simple model to understand accumulation of CO_2 in the atmosphere

user. Anthropogenic emissions are set to match Webster's assumption. They grow to 19 GT/year by the end of the century. Net removal to oceans is assumed to remain constant at 2 GT/year for the reasons given previously. Net removal to biomass and soils is then subject to experimentation to allow this simple model to match Webster's results. A close match is provided if the net removal increases from 1 to 2 GT/year during the first half of the century and then remains at 2 GT/year for the next fifty years. With these assumptions, the CO_2 in the atmosphere would double from 750 to 1500 GT during the century. This means that the atmospheric concentration would double from 350 to 700 ppm, the same result published by Webster [23].

The model in Fig. 13 is no more than an accumulator. This is the simplest of possible models to add insight on the dynamics of CO_2 accumulation in the atmosphere. It includes a single stock and only three flows, with all of the flows specified by the user. There are no feedback relationships which are normally at the core of system dynamics models. This extreme simplification is intended to make the point that simple models may provide perspective on the dynamics of a system. In this case, a simple accumulator can teach one about the sluggish response of atmospheric CO_2 in the wake of reductions in the anthropogenic emissions. As an example, suppose carbon policies were to succeed in cutting global emissions dramatically in the year 2050. By this year, emissions would have reached 10 GT/yr, so the supposed policy would reduce emissions to 5 GT/yr. What might then happen to CO_2 concentrations in the atmosphere for the remainder of the century? Experiments with highly educated adults [21] suggest that some subjects would answer this question with "pattern matching" reasoning. For example, if emissions are cut in half, it might make sense that CO_2 concentrations would be cut in half as well. But pattern matching leads one astray since the accumulation of CO_2 in the atmosphere responds to the total effect of the flows in Fig. 13. Were anthropogenic emissions to be reduced to 5 GT/year and net removals were to remain at 4 GT/year, the CO_2 concentration would continue to grow, and atmospheric CO_2 would reach 470 ppm by the end of the century.

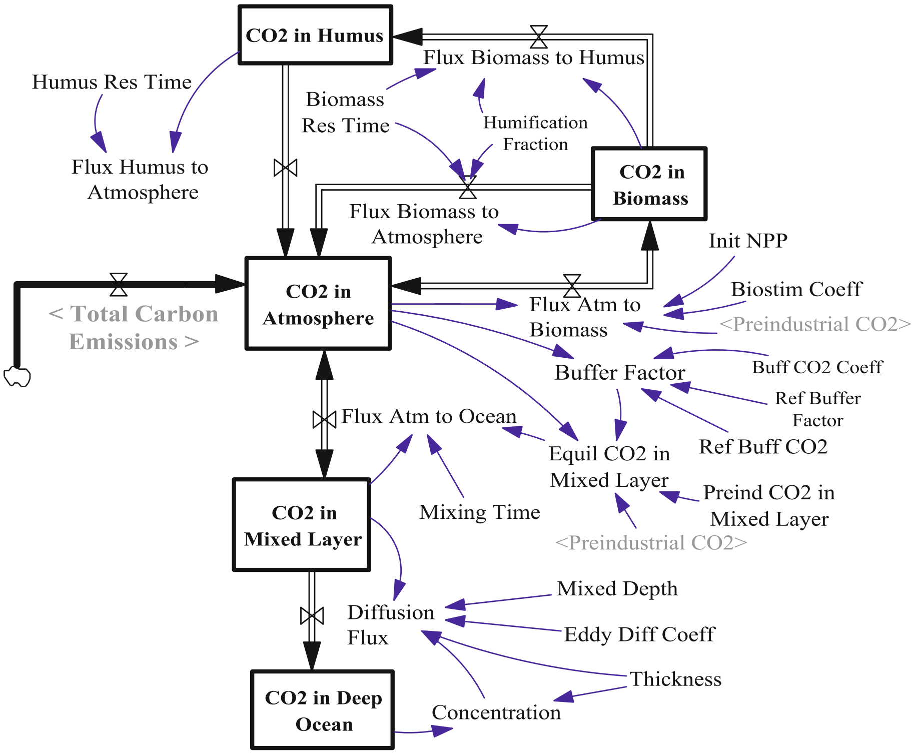
The model in Fig. 13 is an extreme example to make a point about the usefulness of simple models. The next example is by Fiddaman [6]. It was selected as illustrative of the type of model that would emerge after a system dynamics study. Figure 14 shows the view of the carbon cycle, one of 30 views in the model. The model simulates the climate system within a larger system that includes growth in human population, growth in the economy, and changes in the production of energy. The model was organized conceptually as nine interacting sectors with a high degree of coupling between the energy, economic and the climate sectors.

Fiddaman focused on policy making, particularly the best way to put a price on carbon. In the current debate, this question comes down to a choice between a carbon tax and a carbon market. His simulations add support to those who argue that the carbon tax is the preferred method of putting a price on carbon. The simulations also provide another example of the usefulness of system dynamics models that cross disciplinary boundaries. By representing the economy, the energy system and the climate system within a single, tightly coupled model, he provides another example of the power of system dynamics to promote interdisciplinary exploration of complex problems.

System dynamics has also been applied to a wide variety of energy problems [1,7]. Indeed, a key word frequency count in 2004 revealed nearly 400 energy entries in the System dynamics bibliography [11]. Many of these applications deal with the electric power industry, and I have selected two electric studies to illustrate the usefulness of the approach. The first involves the regulatory and financial challenges of the investor owned electric utilities in the United States.

Lessons from the Regulated Power Industry in the 1970s

The 1970s was a difficult decade for the regulated power companies in the United States. The price of oil and gas was increasing rapidly, and the power companies were frequently calling on their regulators to increase retail rates to cover the growing cost of fuel. The demand for electricity had been growing rapidly during previous decades, often at 7 %/year. At this rate, the demand doubled every decade, and the power companies faced the challenge of doubling the amount of generating capacity to ensure that demand would be satisfied. The power companies dealt with this challenge in previous decades by building ever



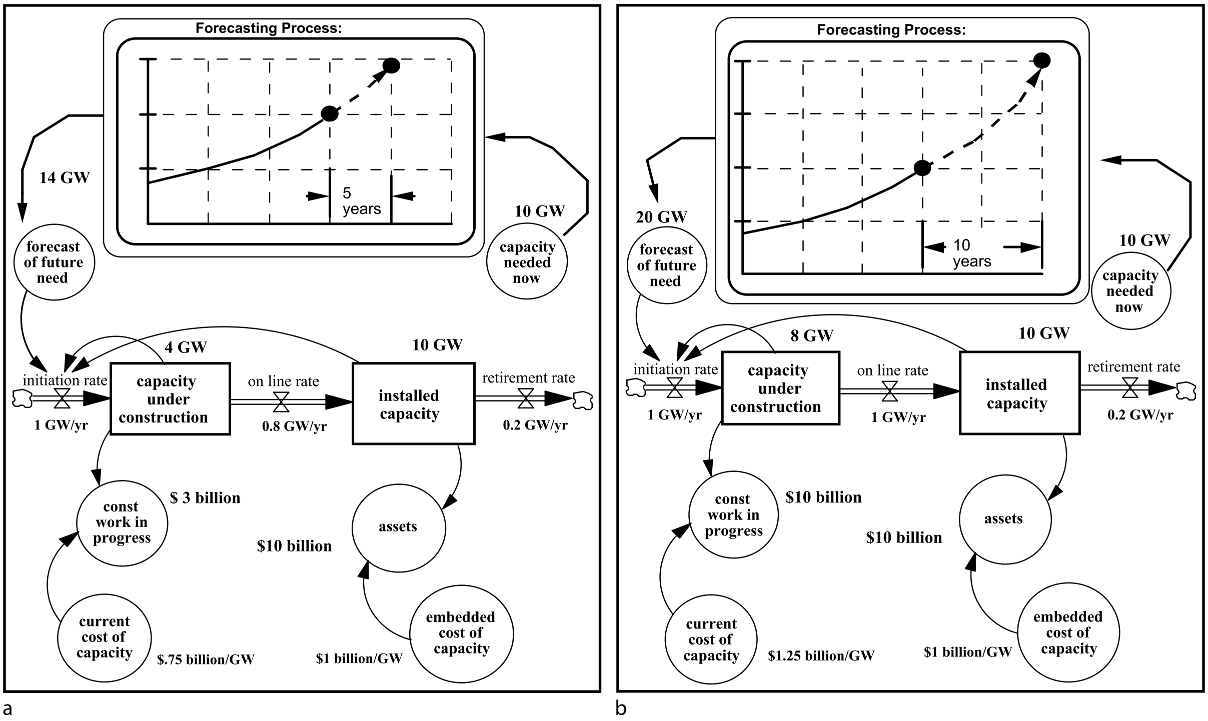
System Dynamics Models of Environment, Energy and Climate Change, Figure 14
Representation of the carbon cycle in the model by Fiddaman [6]

larger power plants (whose unit construction costs declined due to economies of scale). But the economies of scale were exhausted by the 1970s, and the power companies found themselves with less internal funds and poor financial indicators. Utilities worried that the construction of new power plants would not keep pace with demand, and the newspapers warned of curtailments and blackouts.

Figure 15 puts the financial problems in perspective by showing the forecasting, planning and construction processes. The side by side charts allows one to compare the difficult conditions of the 1970s with conditions in previous decades. Figure 15a shows the situation in the 1950s and 1960s. Construction lead times were around 5 years, so forecasts would extend 5 years into the future. Given the costs at the time, the power company would need to finance $3 billion in construction. This was a substantial, but manageable task for a company with $10 billion in assets.

Figure 15b shows the dramatic change in the 1970s. Construction lead times had grown to around 10 years, and construction costs had increased as well. The power company faced the challenge of financing $10 billion in construction with an asset base of $10 billion. The utility executives turned to the regulators for help. They asked for higher electricity rates in order to increase annual revenues and improve their ability to attract external financing. The regulators responded with substantial rate increases, but they began to wonder whether further rate increases would pose a problem with consumer demand. If consumers were to lower electricity consumption, the utility would have less sales and less revenues. The executives might then be forced to request another round of rate increases. Regulators wondered if they were setting loose a "death spiral" of ever increasing rates, declining sales and inadequate financing.

Figure 16 puts the problem in perspective by showing the consumer response to higher electricity rates along side of the other key feedback loops in the system. Higher electricity rates do pose the problem which came to be called "the death spiral". But the death spiral does not act in isolation. Figure 16 reminds us that higher rates lead to lower consumption and to a subsequent reduction in the demand forecast and in construction. After delays for the new power plants to come on line, the power companies experiences a reduction in its "rate base" and the "allowed revenues". When the causal relationships are traced around the outer loop, one sees a negative feedback loop that could act to stabilize the situation. The prob-



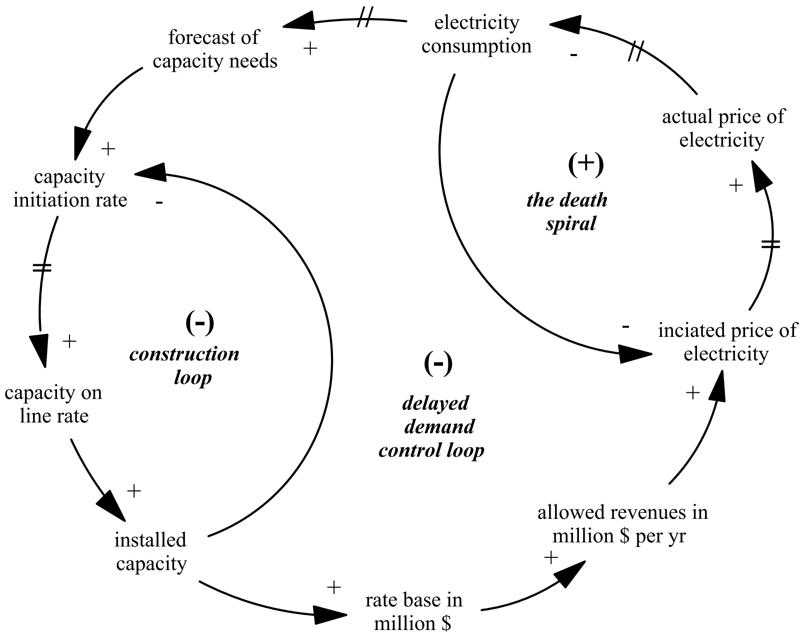
System Dynamics Models of Environment, Energy and Climate Change, Figure 15
a The electric utility's financial challenge during the 1950s and 1960s. b The electric utility's financial challenge during the 1970s

lem, however, is that the delays around the outer loop are substantially longer than the delay for the death spiral.

The utility companies financial challenge was the subject of several system dynamics studies in the 1970s and 1980s [7]. The studies revealed that the downward spiral could pose difficult problems, especially if consumers reacted quickly while utilities were stuck with long-lead time, capital intensive power plants under construction. The studies showed that utility executives needed to do more than rely on regulators to grant rate increases; they needed to take steps on their own to soften the impact of the death spiral. The best strategy was to shift the investments to technologies with shorter lead times. (As an example, a power company in coal region would do better to switch from large to smaller coal plants because of the small plants' shorter lead time.) The studies also revealed that the company's financial situation would improve markedly with slower growth in demand. By the late 1970s and early 1980s, many power companies began to provide direct financial incentives to their customers to slow the growth in demand. System dynamics studies showed that the company-sponsored efficiency programs would be beneficial to the both the customers (lower electric bills) and to the power companies (improved financial performance).

An essential feature of the utility modeling was the inclusion of power operations along side of consumer behavior, company forecasting, power plant construction, regulatory decision making and company financing. This interdisciplinary approach is common within the system dynamics community because practitioners believe that insights will emerge from simulating the key feedback loops. (This belief leads one to follow the cause and effect connections around the key loops regardless of the disciplinary boundaries that are crossed along the way.) This approach contrasts strongly with the customary modeling framework of large power companies who were not familiar with system dynamics. Their approach was to assign models to different departments (i. e., operations, accounting and forecasting) and string the models together to provide a view of the entire corporation over the long-term planning interval.

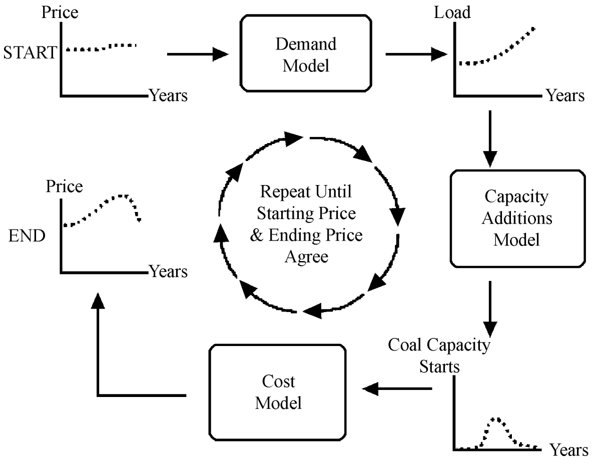
Figure 17 shows what can happen when models within separate departments are strung together. A large corporation might use 30 models, but this diagram makes the point by describing three models. The analysis would begin with an assumption on future electricity prices over the



System Dynamics Models of Environment, Energy and Climate Change, Figure 16
Key feedbacks and delays faced by power companies in the 1970s

20-year interval. These are needed to prepare a forecast of the growth in electricity load. The forecast is then given to the planning department which may run a variety of models to select the number power plants to construct in the future. The construction results are then handed to the accounting and rate making departments to prepare a forecast of electricity prices. When the company finally completes the many calculations, the prices that emerge may not agree with the prices that were assumed at the start. The company must then choose whether to ignore the contradiction or to repeat the entire process with a new estimate of the prices at the top of the diagram. This was not an easy choice. Ignoring the price discrepancy was problematic because it was equivalent to ignoring the "death spiral," one of the foremost problems of the 1970s. Repeating the analysis was also problematic. The new round of calculations would be time consuming, and there was no guarantee that consistent results would be obtained at the end of the next iteration.

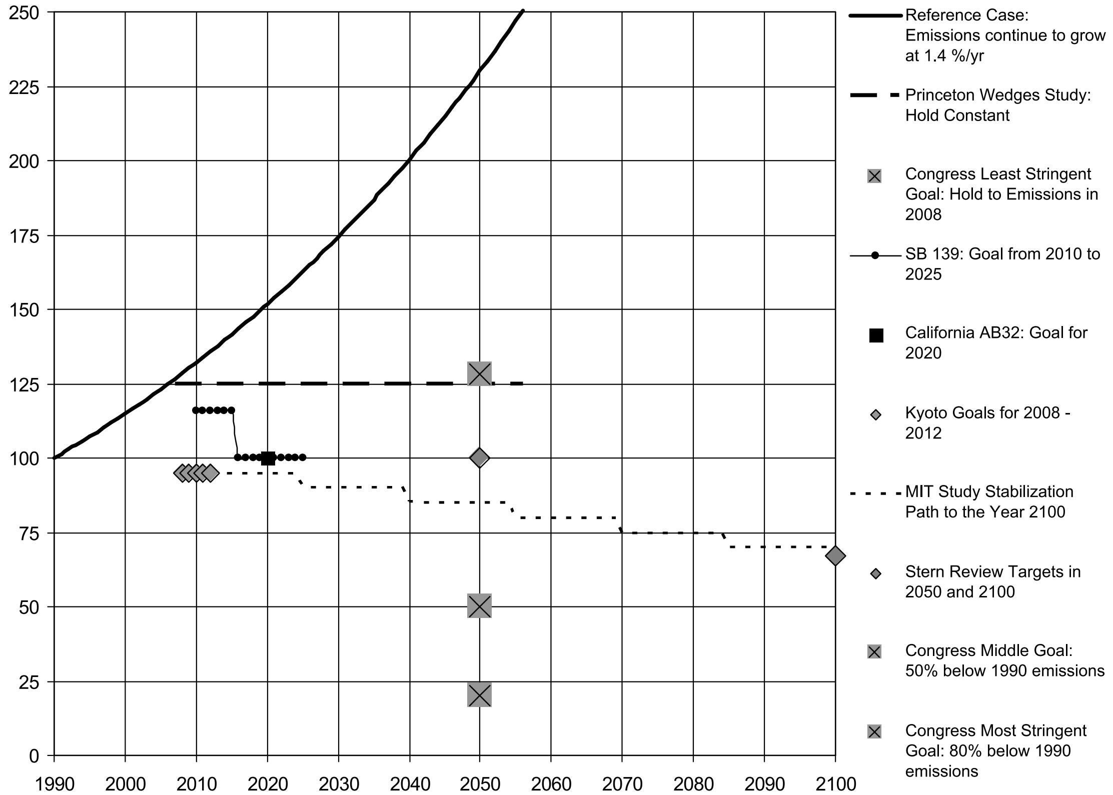
The power companies' dilemma from the 1970s is described here to make an important point about the usefulness of system dynamics. System dynamics modeling is ideally suited for the analysis of dynamic problems that require a feedback perspective. The method allows one to "close the loop", as long as one is willing to cross the necessarily disciplinary boundaries. In contrast, other modeling methods are likely to be extremely time consuming or fall short in simulating the key feedbacks that tie the system together.



System Dynamics Models of Environment, Energy and Climate Change, Figure 17
The iterative approach often used by large power companies in the 1970s

Simulating the Power Industry Response to a Carbon Market

The world is getting warmer, both in the atmosphere and in the oceans. The clearest and most emphatic description

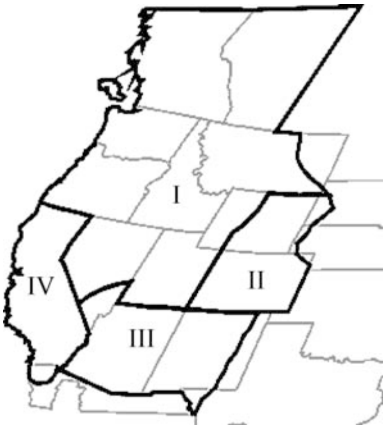


System Dynamics Models of Environment, Energy and Climate Change, Figure 18
Comparison of goals for emissions (100 on the vertical axis represent emissions in the year 1990)

of global warming was issued by the intergovernmental panel on climate change (IPCC) in February of 2007. Their summary for policymakers (p. 4 in [14]) reported that the "Warming of the climate system is unequivocal, as is now evident from observations of increases in global average air and ocean temperatures, widespread melting of snow and ice and rising global mean sea level". The IPCC concluded that "most of the observed increase is very likely due to the observed increase in anthropogenic greenhouse gas concentrations". As a consequence of the IPCC and other warnings, policymakers around the world are calling for massive reductions in CO_2 and other greenhouse gas (GHG) emissions to reduce the risks of global warming.

Figure 18 summarizes some of the targets for emission reductions that have been adopted or proposed around the world. In many cases, the targets are specified relative to a country's emissions in the year 1990. So, for ease of comparison, the chart uses 100 to denote emissions in the year 1990. Emissions have been growing at around 1.4%/year. The upward curve shows the future emissions if this trend continues: emissions would reach 200 by 2040 and 400 by 2090. The chart shows the great differences in the stringency of the targets. Some call for holding emissions constant; others call for dramatic reductions over time. Some targets apply to the next two decades; many extend to the year 2050; and some extend to the year 2100. However, when compared to the upward trend, all targets require major reductions relative to business as usual.

The targets from the Kyoto treaty are probably the best known of the goals in Fig. 18. The treaty became effective in February of 2005 and called for the Annex I countries to reduce emissions, on average, by 5% below 1990 emissions by the year 2008 and to maintain this limit through 2012. The extension of the Kyoto protocol beyond 2012 is the subject of ongoing discussions. The solid line from 2010 to 2050 represents the "stabilization path" used in the climate modeling by Webster [23]. The limit on emissions was imposed in modeling calculations designed to stabilize atmospheric CO_2 at 550 ppmv or lower. The scenario assumed that the Kyoto emissions caps are adopted by all countries by 2010. The policy assumed that the caps would be extended and then further lowered by 5% every 15 years. By the end of the century, the emissions would be 35% below the value in 1990.



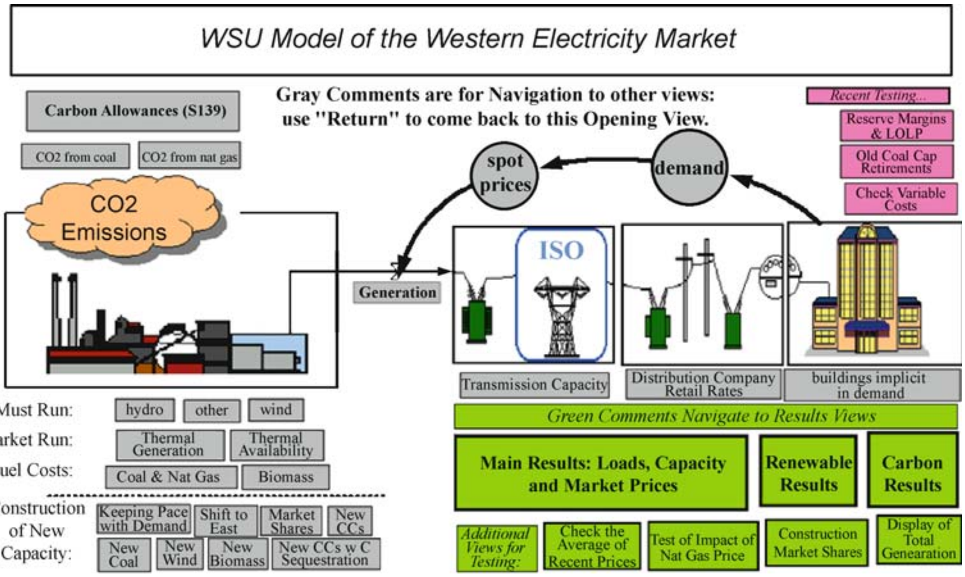
System Dynamics Models of Environment, Energy and Climate Change, Figure 19
Map of the western electricity system

This article concentrates on Senate Bill 139, The Climate Stewardship Act of 2003. Figure 19 shows the S139 targets over the interval from 2010 to 2025. The bill called for an initial cap on emissions from 2010 to 2016. The cap would be reduced to a more challenging level in 2016, when the goal was to limit emissions to no more than the emissions from 1990. S139 was introduced by Senators McCain and Lieberman in January of 2003. It did not pass, but it was the subject of several studies including a highly detailed study by the Energy Information Administration [5]. The EIA used a wide variety of models to search for the carbon market prices that would induce industries to lower emissions to come into compliance with the cap. The carbon prices were estimated at $22 per metric ton of CO_2 when the market was to open in 2010. They were projected to grow to $60 by the year 2025.

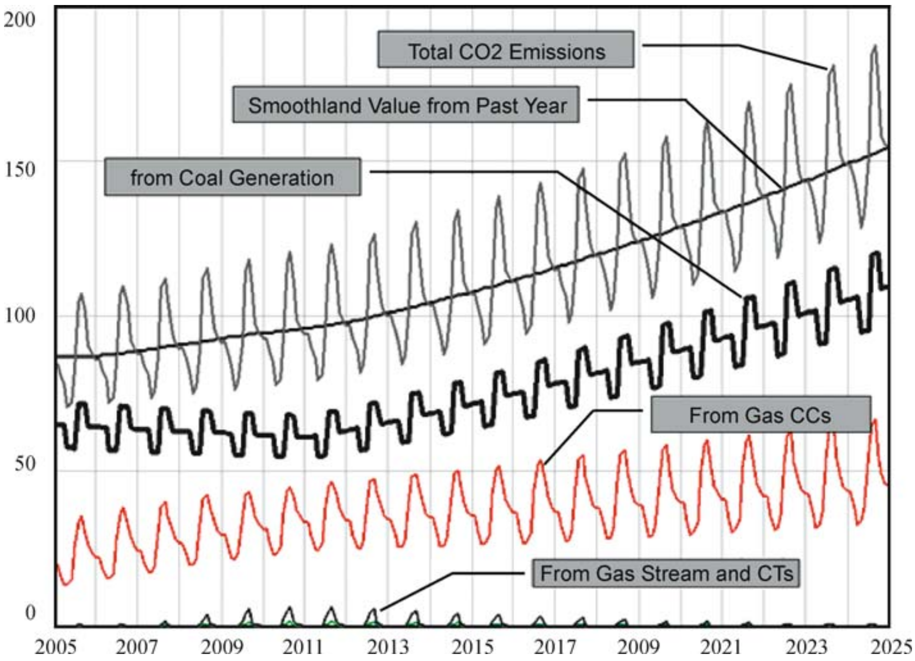
The EIA study showed that the electric power sector would lead the way in reducing emissions. By the year 2025, power sector emissions would be reduced 75% below the reference case. This reduction was far beyond the reductions to be achieved by other sectors of the economy. This dramatic response was possible given the large use of coal in power generation and the power industry's wide range of choices for cleaner generation.

A system dynamics study of S139 was conducted at Washington State University (WSU) to learn if S139 could lead to similar reductions in the west. Electricity generation in the western system is provided in a large, interconnected power system shown in Fig. 19. This region has considerably more hydro resources, and it makes less use of coal-fired generation than the nation as a whole. The goal was to learn if dramatic reductions in CO_2 emissions could be possible in the west and to learn if they could be achieved with generating technologies that are commercially available today.

The opening view of the WSU model is shown in Fig. 20. The model deals with generation, transmission and distribution to end use customers, with price feedback on the demand for electricity. The model is much larger than the textbook models described earlier in this article. Fifty


WSU Model of the Western Electricity Market
Carbon Allowances (S139)
CO2 from coal
CO2 from nat gas
Gray Comments are for Navigation to other views:
use "Return" to come back to this Opening View.
Recent Testing...
Reserve Margins & LOLP
Old Coal Cap Retirements
Check Variable Costs
CO2 Emissions
spot prices
demand
Generation
ISO
Transmission Capacity
Distribution Company Retail Rates
buildings implicit in demand
Green Comments Navigate to Results Views
Must Run:
hydro
other
wind
Market Run:
Thermal Generation
Thermal Availability
Fuel Costs:
Coal & Nat Gas
Biomass
Construction of New Capacity:
Keeping Pace with Demand
Shift to East
Market Shares
New CCs
New Coal
New Wind
New Biomass
New CCs w C Sequestration
Main Results: Loads, Capacity and Market Prices
Renewable Results
Carbon Results
Additional Views for Testing:
Check the Average of Recent Prices
Test of Impact of Nat Gas Price
Construction Market Shares
Display of Total Genearation

System Dynamics Models of Environment, Energy and Climate Change, Figure 20
Opening view of the model of the western electricity system



System Dynamics Models of Environment, Energy and Climate Change, Figure 21
Annual emissions in a base case simulation (annual emissions are in million metric tons of carbon)

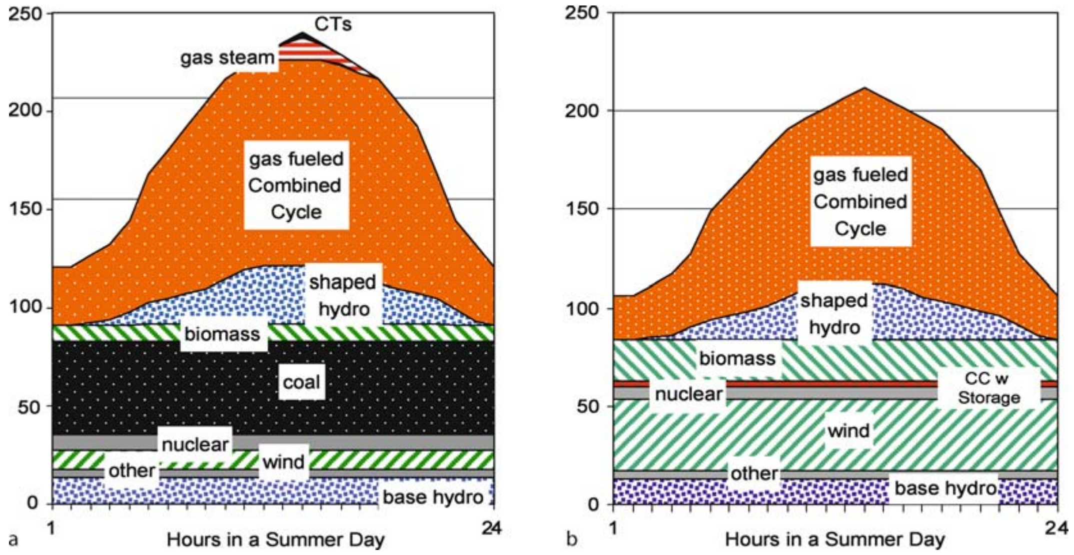
views are required to show the all the diagrams and the simulation results. The opening view serves as a central hub to connect with all the other views.

The opening view uses Vensim's comment icons to draw attention to the CO_2 emissions in the model. The emissions arise mainly from coal-fired power plants, as shown in Fig. 21. A smaller, but still significant fraction of the emissions is caused by burning natural gas in combined cycle power plants. Total emissions vary with the seasons of the year, with the peak normally appearing in the summer when almost all of the fossil-fueled plants are needed to satisfy peak demand. The base case shows annual emissions growing by over 75% by the year 2025.

A major challenge for the system dynamics model is representing power flows across a transmission grid. Finding the flows on each transmission line and the prices in each area is difficult with the standard tools of system dynamics. It simply doesn't make sense to represent the power flows with a combination of stocks, flows and feedback processes to explain the flows. It makes more sense to calculate the flows and prices using traditional power systems methods, as explained by Dimitrovski [4]. The power flows were estimated using an algebraic approach which power engineers label as a reduced version of a direct-current optimal power flow calculation. The solution to the algebraic constraints were developed with the Matlab software and then transferred to user-defined functions to operate within the Vensim software. The Vensim simulations were set to run over twenty years with time in months. (A typical simulation required 240 months with changes during a typical day handled by carrying along separate calculations for each of 24 h in a typical day.) These are extensive calculations compared to many system dynamics models, so there was concern that we would lose the rapid simulation speed that helps to promote interactive exploration and model testing. The important methodological accomplishment of this project was the inclusion of network and hourly results within a long-term model without losing the rapid simulation response that encourages users to experiment with the model.

One of the model experiments called for a new simulation with carbon prices set to follow the $20 to $60 trajectory projected by the EIA for S139. These prices were specified as a user input, and the model responded with a change in both short-term operations and long-term investments. The important result was a 75% reduction in CO_2 emissions by the end of the simulation. This dramatic reduction corresponds almost exactly to the EIA estimate of CO_2 reduction for the power industry in the entire US.

Figure 22 helps one understand how CO_2 emissions could be reduced by such a large amount. These diagrams show the operation of generating units across the Western US and Canada for a typical day in the summer of the final year of the simulation. Figure 22a shows the reference



System Dynamics Models of Environment, Energy and Climate Change, Figure 22

a Projected generation for a peak summer day in 2024 in the reference case. b Projected generation for a peak summer day in 2024 in the S139 case

case; Figure 22b shows the case with S139. The side by side comparison helps one visualize the change in system operation. A comparison of the peak loads shows that the demand for electricity would be reduced. The reduction is 9%, which is due entirely to the consumers' reaction to higher retail electric prices over time.

Figure 22b shows large contributions from wind and biomass generation. Wind generation is carbon free, and biomass generation is judged to be carbon neutral, so these generating units make an important contribution by the end of the simulation. Both of these generating technologies are competitive with today's fuel prices and tax credits. The model includes combined cycle gas generation equipped with carbon capture and storage, a technology that is not commercially available today. The model assumes that advances in carbon sequestration over the next two decades would allow this technology to capture a small share of investment near the end of the simulation. By the year 2025, the combined cycle plants with sequestration equipment would provide 2% of the generation.

The most important observation from Fig. 22 is the complete elimination of coal-fired generation in the S139 case. Coal-fired units are shown to operate in a base load mode in the reference simulation. They provide around 28% of the annual generation, but they account for around two/thirds of the CO_2 emissions in the western system. The carbon prices from S139 make investment in new coal-fired capacity unprofitable at the very start of the simulated market in 2010. As the carbon prices increase, utilities to cut back on coal-fired generation and compensate with increased generation from gas-fired CC capacity. In the simulations reported here, this fuel switching would push the coal units into the difficult position of operating fewer and fewer hours in a day. Eventually this short duration operation is no longer feasible, and coal generation is eliminated completely by the end of the simulation.

The WSU study of the western electric system was selected as the concluding example because of its novel treatment of network flows inside a system dynamics model [4]. The model is also interesting for its treatment of daily price changes within a long-term model. (Such changes are important to the simulation of revenues in the wholesale market.) From a policy perspective, the study confirms previous modeling of the pivotal role of the electric power industry in responding to carbon markets. The study indicated that the western electricity system could achieve dramatic reductions in CO_2 emissions within 15 years after the opening of a carbon market, and it could do so with technologies that are commercially available today [8].

Conditions for Effective Interdisciplinary Modeling

All of the applications demonstrate the usefulness of system dynamics in promoting interdisciplinary modeling. The article concludes with comments on the level of effort and the conditions needed for effective, interdisciplinary modeling.

The examples in this article differ substantially in the level of effort required, from several weeks for the classroom examples to several years for the energy studies. The textbook examples involved student expansions of models of Mono Lake and the Tucannon salmon. The expansions were completed by undergraduate students in projects lasting two or three weeks. The key was the students' previous education (classes from many different departments) and their receptiveness to an interdisciplinary approach.

Fiddaman's model of the climate and energy system [6] was a more ambitious exercise, requiring several years of effort as part of his doctoral research. Bringing multi-year interdisciplinary modeling projects to a successful conclusion requires one to invest the time to master several disciplines and to maintain a belief that there are potential insights at the end of the effort.

The electric power industry examples were also ambitious projects that required several years of effort. The modeling of the western electricity system was a four-year project with support from the National Science Foundation. The long research period was crucial for it allowed the researchers from power systems engineering, system dynamics and environmental science to take the time to learn from one another. The modeling of the electric company problems in the 1970s was also spread over several years of effort. The success of this modeling was aided by utility planners, managers and modelers who were looking for a systems view of their agency and its problems. They saw system dynamics as a way to tie existing ideas together within an integrated portrayal of their system. Their existing ideas were implemented in models maintained by separate functional areas (i. e., forecasting, accounting, operations). The existing models often provided a foundation for the system dynamics models (i. e., in the same way that the comprehensive climate models in Fig. 11 provide support for the development of the more integrated models). The key to effective, interdisciplinary modeling within such large organizations is support from a client with a strong interest in learning and with managerial responsibility for the larger system.

Future Directions

This article concludes with future directions for system dynamics applications to climate change. People often talk of mitigation and adaptation. Mitigation refers to the challenge of lowering greenhouse gas emissions to avoid dangerous anthropogenic interference with the climate system. Adaptation refers to the challenge of living in a changing world.

Mitigation: The challenge of lowering CO_2 and other GHG emissions is the fundamental challenge of the coming century. The next two decades will probably see various forms of carbon markets, and system dynamics can aid in learning about market design. It is important that we learn how to make these markets work well. And if they don't work well, it's important to speed the transition to a carbon tax policy with better prospects for success. System dynamics can aid in learning about markets, especially if it is coupled with simulating gaming to allow market participants and regulators to "experience" and better understand market dynamics.

Adaptation: The world will continue to warm, and sea levels will continue to rise. These trends will dominate the first half of this century even with major reductions in CO_2 emissions. These and other climate changes will bring a wide variety of problems for management of water resource, public health planning, control of invasive species, preservation of endangered species, control of wildfire, and coastal zone management, just to name a few. Our understanding of the adaptation challenges can be improved through system dynamics modeling. The prospects for insight are best if the models provide an interdisciplinary perspective on adapting to a changing world.

Bibliography

Primary Literature

1. Bunn D, Larsen E (1997) Systems modelling for energy policy. Wiley, Chichester
2. Claussen M et al (2002) Earth system models of intermediate complexity: closing the gap in the spectrum of climate system models. Climate Dyn 18:579–586
3. Coyle G (1977) Management system dynamics. Wiley, Chichester
4. Dimitrovski A, Ford A, Tomsovic K (2007) An interdisciplinary approach to long term modeling for power system expansion. Int J Crit Infrastruct 3(1–2):235–264
5. EIA (2003) United States Department of Energy, Energy Information Administration, Analysis of S139, the Climate Stewardship Act of 2003
6. Fiddaman T (2002) Exploring policy options with a behavioral climate-economy model. Syst Dyn Rev 18(2):243–264
7. Ford A (1999) Modeling the environment. Island Press, Washington
8. Ford A (2008) Simulation scenarios for rapid reduction in carbon dioxide emissions in the western electricity system. Energy Policy 36:443–455
9. Forrester J (1961) Industrial dynamics. Pegasus Communications
10. Forrester J (2000) From the ranch to system dynamics: An autobiography, in management laureates. JAI Press
11. Ford A, Cavana R (eds) (2004) Special Issue of the Syst Dyn Rev
12. Hardin G (1968) The tragedy of the commons. Science 162:1243–1248

13. IPCC (1997) An introduction to simple climate models used in the IPCC second assessment report. ISBN 92-9169-101-1
14. IPCC (2007) Climate change 2007: The physical science basis, summary for policymakers. www.ipcc.ch/
15. Kump L (2002) Reducing uncertainty about carbon dioxide as a climate driver. Nature 419:188–190
16. Meadows DH, Meadows DL, Randers J, Behrens W (1972) The limits to growth. Universe Books
17. Morecroft J (2007) Strategic modelling and business dynamics. Wiley, Chichester
18. Richardson J, Pugh A (1981) Introduction to system dynamics modeling with dynamo. Pegasus Communications
19. Sterman J (2000) Business dynamics. McGraw-Hill, Irwin
20. Sterman J (ed) (2002) Special Issue of the Syst Dyn Rev
21. Sterman J, Sweeney L (2007) Understanding public complacency about climate change. Clim Chang 80(3–4):213–238
22. Warren K (2002) Competitive strategy dynamics. Wiley, Chichester
23. Webster M et al (2003) Uncertainty analysis of climate change and policy response. Climat Chang 61:295–320

Books and Review

Houghton J (2004) Global warming: The complete briefing, 3rd edn. Cambridge University Press, Cambridge

System Dynamics Models, Optimization of

BRIAN DANGERFIELD
Centre for OR & Applied Statistics, Salford Business School, University of Salford, Salford, UK

Article Outline

Glossary

Econometrics A statistical approach to economic modeling in which all the parameters in the structural equations are estimated according to a 'best fit' to historical data.

Maximum likelihood A statistical concept which underpins calibration optimization and which generates the most likely parameter values; it is equivalent to the parameter set which minimizes the chi-square value.

Objective function See **Payoff** below.

Optimization The process of improving a model's results in terms of either an aspect of its performance or by calibrating it to fit reported time series data.

Payoff A formula which expresses the objective, say, maximization of profits, minimization of costs or minimization of the differences between a model variable and historical data on that variable.

Zero-one parameter A parameter which is used as a multiplier in a policy equation and serves the effect of bringing in or removing a particular influence in determining the optimal policy.

Definition of the Subject

The term 'optimization' when related to system dynamics (SD) models has a special significance. It relates to the mechanism used to improve the model vis-à-vis a criterion. This collapses into two fundamentally different intentions. Firstly one may wish to improve the model in terms of its performance. For instance, it may be desired to minimize overall costs of inventory whilst still offering a satisfactory level of service to the downstream customer. So the criterion here is cost, and this would be minimized after searching the parameter space related to service level. The direction of need may be reversed and maximization may be desired as, for instance, if one had a model of a firm and wished to maximize profit subject to an acceptable level of payroll and advertising costs. Here the parameter space being explored would involve both payroll and advertising parameters. This type of optimization might be described generically as *policy optimization*.

Optimization of performance is also the *raison d'etre* of other management science tools, most notably mathematical programming. But such tools are usually static: they offer the 'optimum' resource allocation given a set of constraints and a performance function to either maximize or minimize. These models normally relate to a single time point and may then need to be re-run on a weekly or monthly basis to determine a new optimal resource allocation. In addition, these models are often linear (certainly so in the case of linear programming), whereas SD models are usually non-linear. So the essential differences are that SD model optimization for performance involves both a dynamic and a non-linear model.

A separate improvement to the model may be sought where it is required to fit the model to past time series data. Optimization here involves minimizing a statistical function which expresses how well the model fits a time-series of data pertaining to an important model variable. In other words a vector of parameters are explored with a view to determining the particular parameter combination which offers the best fit between the chosen important model variable and a past time series data set of this variable. This type of optimization might be generically termed *model calibration*. If *all* the parameters in the SD model are determined in this fashion then the process is equivalent to the technique of econometric modeling. A good comparison between system dynamics and econometric modeling can be found in Meadows and Robinson [12].

Optimization as Calibration

In these circumstances we wish to determine optimal parameters, those which, following a search of the parameter space, offer the best fit of a particular model variable to a time series dataset on that variable taken from real world reporting.

As an example consider a variation of the one of the epidemic models which are made available with the Vensim™ software. The stock-flow diagram is presented as Fig. 1.

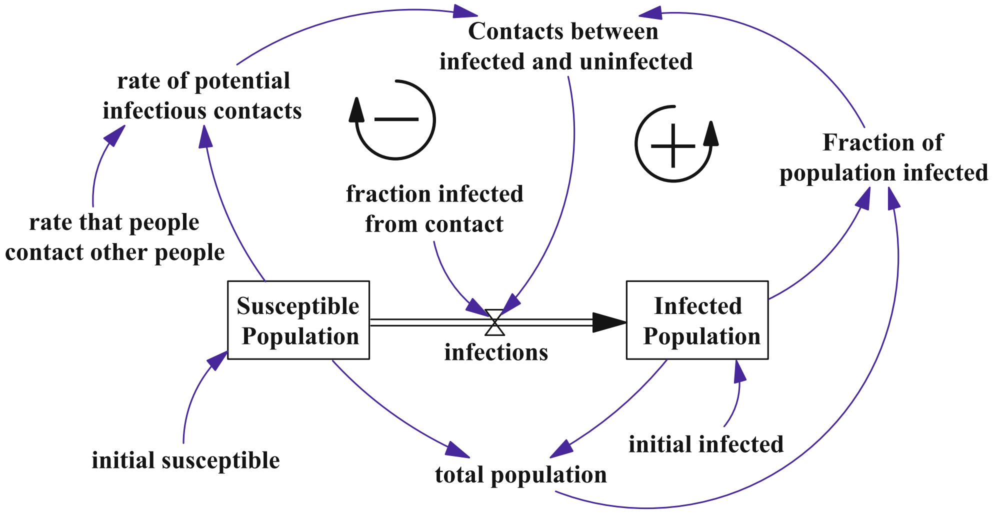
In this epidemiological system members of a susceptible population become infected and join the infected population. Epidemiologists call this an S–I model. It is a simpler variation of the S–I–R model which includes recovered (R) individuals.

Suppose some data on new infections (at intervals of five days) are available covering 25 days of a real-world epidemic. The model is set with a time horizon of 50 days which is consistent with, say, a flu epidemic or an infectious outbreak of dysentery in a closed population such as a cruise ship. The 'current' run of the model is shown in Fig. 2, with the real-world data included for comparison.

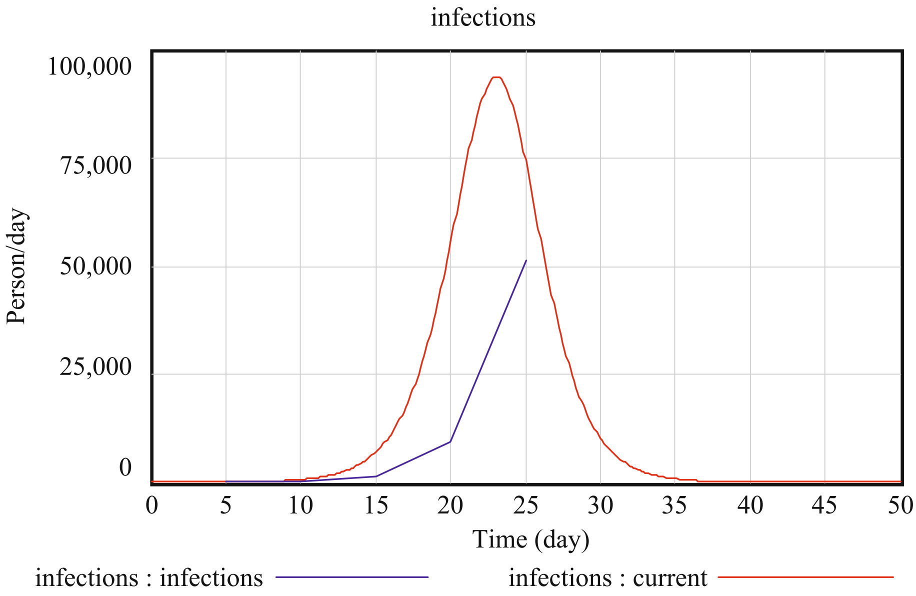
Clearly there is not a very good correspondence between the actual data and the model variable for the infection rate (infections). We wish to achieve a better calibration, and so there is a need to select relevant parameters through which the calibration optimization can be performed over. Referring back to Fig. 1, we can see that the *fraction infected from contact* and the *rate that people contact other people* are two possible parameters to consider. The initial infected and initial susceptible are also parameters of the model in the strict sense of the term, but we will ignore them on this occasion. In this model the *initial infected* is 10 persons and *initial susceptibles* number 750,000 persons.

The chosen value for the *fraction infected from contact* is 0.1, while that for the *rate that people contact other people* is 5.0. The former is a dimensionless number while the latter is measured as a fraction per day (1/day). This is obtained from consideration of the *rate of potential infectious contacts* (persons/day) as a proportion of the *susceptible population* (persons).

The optimization process for calibration involves reading into the model the time series data, in this case on new infections, and, secondly, determining the range for the search in parameter space. There is usually some basic background knowledge which allows a sensible range to be entered. For instance, a probability can only be specified between 0 and 1.0. In this case we have chosen to specify



System Dynamics Models, Optimization of, Figure 1
Stock-flow diagram for a simple epidemic model



System Dynamics Models, Optimization of, Figure 2
Current (base) run of the model and reported data on infections

the ranges as follows:

$$0.03 \leq \text{fraction infected from contact} \leq 0.7$$

$$2 \leq \text{rate that people contact other people} \leq 10\,.$$

A word of warning is necessary in respect of optimizing delay parameters. Because there is a risk of mathematical instability in the model if the value of DT (the TIME STEP) is too large relative to the smallest first-order delay constant, it is important to ensure the TIME STEP employed in the model is sufficiently small to cope with delay constant values which may be reached during the search of the delay parameter space. In other words ensure the minimum number for the search range on the delay parameter is at least double the value of the TIME STEP.

Maximum Likelihood Estimation and the Payoff Function

The optimization process involves a determination of what are termed statistically as maximum likelihood estimates. In Vensim™ this is achieved by maximizing a payoff function. Initially this is negative and the optimization process should ensure this becomes less negative. An ideal payoff value, after optimization, would be zero. A weighting is needed in the payoff function too, but for calibration optimization this is normally 1.0. Driving the payoff value to be larger by making it less negative has parallels with the operation with the simplex algorithm common in linear programming. This algorithm was conceived initially for problems where the objective function was to be minimized. Its use on maximization problems is achieved by minimizing the negative of the objective function.

During the calibration search, Vensim™ takes the difference between the model variable and the data value, multiplies it by the weight, squares it and adds it to the error sum. This error sum is minimized. Usually data points will not exist at every time point in the model. Here the model TIME STEP is 0.125 (1/8th), but let us assume that reported data on new infections have been made available only at times $t = 5, 10, 15, 20$ and 25 so the sum of squares operation is performed only at these five time points.

The data are shown as Table 1.

System Dynamics Models, Optimization of, Table 1
Data used for calibration experiment

Time	5	10	15	20	25
Infections	30	230	1400	9500	51 400

The Recording Point for Reported Data

System dynamics models differentiate between stock and flow variables and the software used for simulating such models advances by a small constant TIME STEP (also known as DT). This has implications for the task of fitting real-world reported data to each type of system dynamics model variable. The following is the issue: at what point in a continuum of time steps should the reported data be recorded at? This is important because the reported data has to be read into the model to be compared with the simulated data. The answer will be different for stock and flow variables.

Where the reported data relate to a stock variable the appropriate time point for recording will be known. If it is recorded at the end of the day (say a closing bank balance) then the appropriate point for data entry in the model will be the beginning of the next day. Thus the first data point above is at time $t = 5$ (5.00) and would, if it were a stock, correspond to a record taken at the very end of time period 4.

However, if the data relate to a flow variable, as in the case of new infections here, the number is the total new infections which have occurred over the entire time unit (day, week, month etc.) and so there is a decision to be reached as to which time point the data are entered at. This is because the TIME STEP (DT) is hardly ever as large as the basic time unit which the model is calibrated in. The use of 5 (10, 15 etc.) above implies that the data on new infections over the period of time $t = 0$ to $t = 5$ are compared with the corresponding model variable at time $5 + 1 * \mathrm{DT}$ (and the new infections over the period $t = 5$ to $t = 10$ at time $10 + 1 * \mathrm{DT}$ etc.). A more appropriate selection might be towards the end of the 5-day time period. Following the example above using a TIME STEP $= 0.125$, this might be at time $4 + 7 * \mathrm{DT}$ (that is at 4.875).

Calibration Optimization Results

Based upon the data on new infections shown above and the chosen ranges for the parameter search, the following output is obtained (Table 2). After 114 simulations the optimized values for our two parameters are shown to be 0.08 and 5.12 and the payoff is over 2500 times larger (less negative). Replacing the original parameters with the optimized values reveals the result shown in Fig. 3. To take things further we may wish to put confidence intervals on the estimated parameters. One way of accomplishing this is by profiling the likelihood and is described in Dangerfield and Roberts [3].

System Dynamics Models, Optimization of, Table 2
Results from the calibration optimization

```
Initial point of search
fraction infected from contact = 0.1
rate that people contact other people = 5
Simulations = 1
Pass = 0
Payoff = −2.67655e + 009
---------------------------------------------
Maximum payoff found at:
fraction infected from contact = 0.0794332
*rate that people contact other people = 5.11568
Simulations = 114
Pass = 6
Payoff = −1.06161e + 006
```

Avoid Cumulated Data

There might be a temptation to optimize parameters against cumulated data when the data are reported essentially as a flow, as is the case here. Were the data to be cumulated we would obtain as shown in Table 3.

The results from this optimization are shown in Table 4. The ranges for the parameter space search are kept the same but the payoff function now involves a comparison of the model variable *infected population* with the corresponding cumulated data. Figure 4 shows the resultant fit to infected population is good, but that is manifestly not borne out when we consider the plot of infections obtained from the same optimization run (Fig. 5).

The reason for this is rooted in statistics. The maximum likelihood estimator is equivalent to the chi-squared statistic. This is turn assumes that each expected data value is independent. A cumulated data series would not exhibit this property of independence.

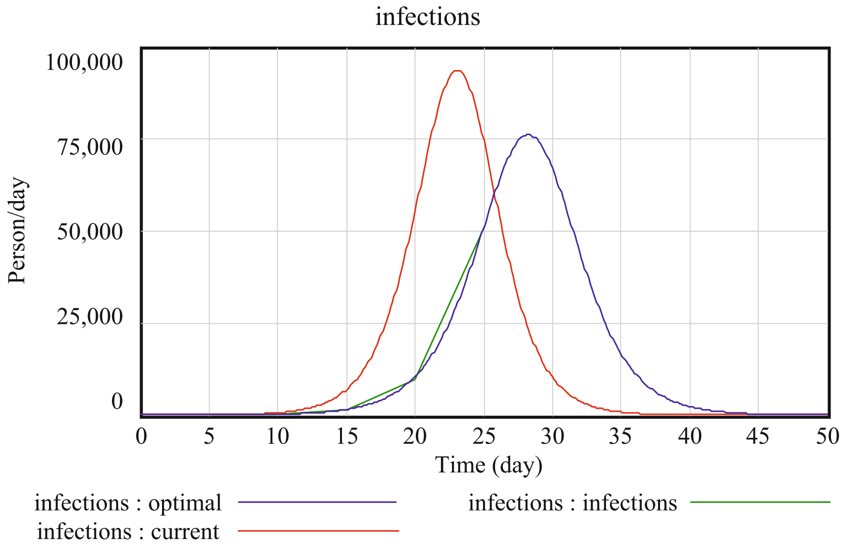
As an aside it is worth pointing out that this model, with suitable changes to the variable names and the time constants involved, could equally represent the diffusion

System Dynamics Models, Optimization of, Table 3
Cumulated reported data for the infected population

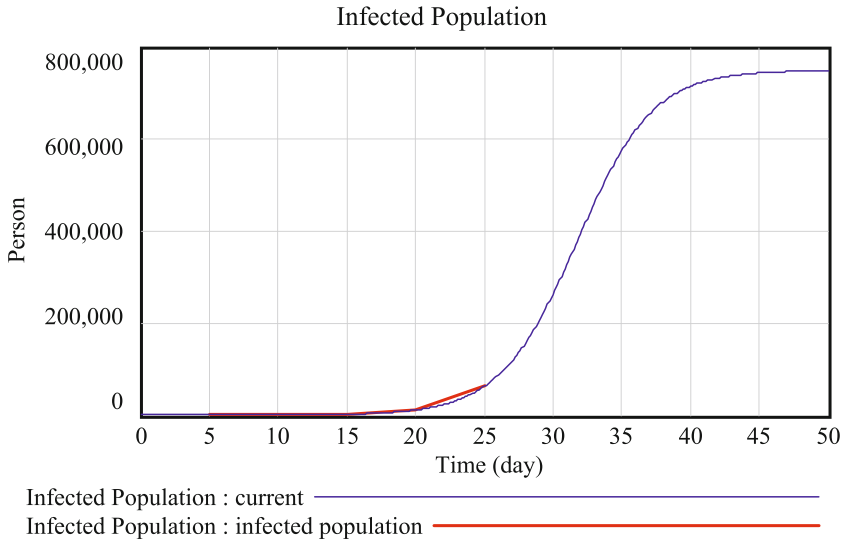
Time	5	10	15	20	25
Infected population	30	260	1660	11 160	62 560

System Dynamics Models, Optimization of, Table 4
Results from the calibration using cumulated data

```
Initial point of search
fraction infected from contact = 0.1
rate that people contact other people = 5
Simulations = 1
Pass = 0
Payoff = −2.48206e + 011
---------------------------------------------
Maximum payoff found at:
fraction infected from contact = 0.0726811
*rate that people contact other people = 4.96546
Simulations = 145
Pass = 6
Payoff = −212 645
The final payoff is −212 645
```



System Dynamics Models, Optimization of, Figure 3
Reported data on infections and optimized (calibrated) model; the base case (current) is reproduced for reference

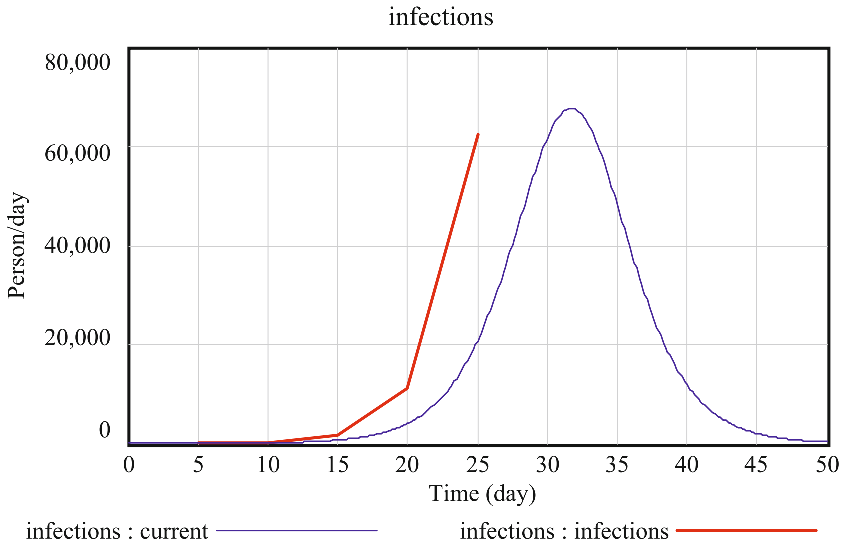


System Dynamics Models, Optimization of, Figure 4
The cumulative model variable (infected population) together with reported data

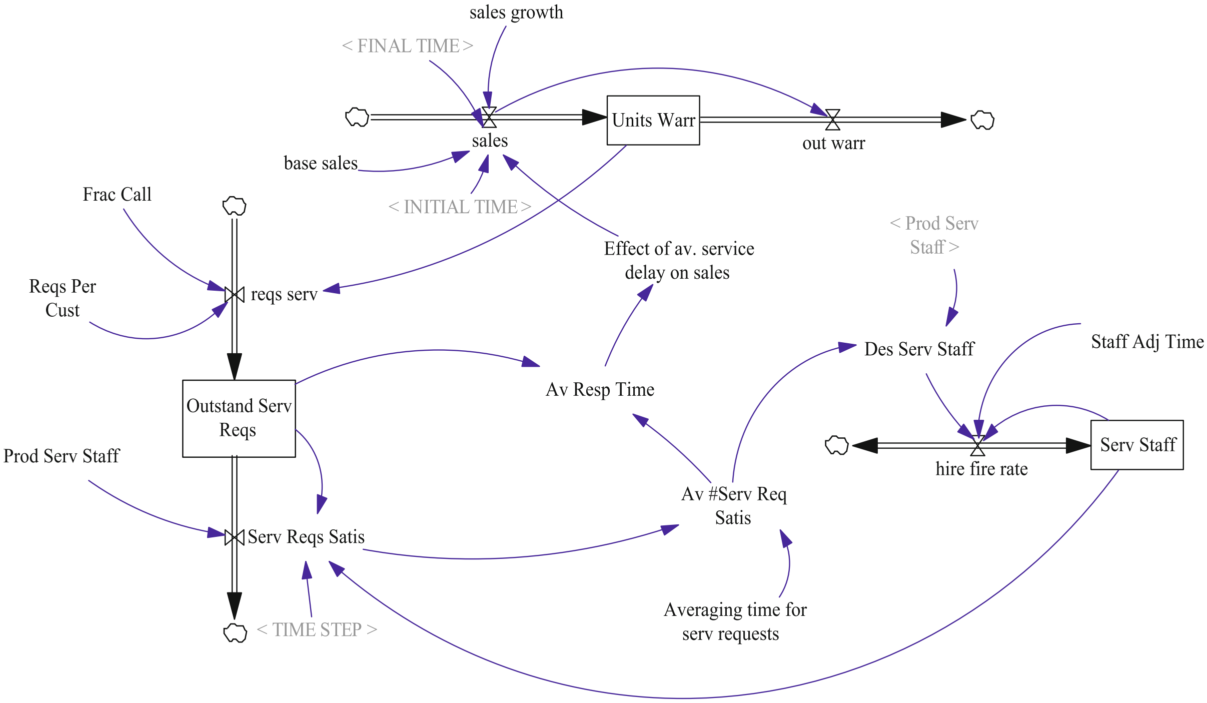
of a new product into a virgin market. In systems terms the structures are equivalent. The *fraction infected from contact* is the same as, say, the fraction reached by word of mouth or advertising and the *rate that people contact other people* is a measure of the potential interactions at which new products might be mentioned amongst the members of the relevant market segment. An *infected population* is equivalent to a customer base, the number of adopters of the relevant product. So it is possible to shed light on important real-world marketing parameters through a calibration optimization of models of this general structure.

Optimization of Performance (Policy Optimization)

An example model is to be used to illustrate the process of optimization to improve the performance of the system, and this is illustrated in Fig. 6. It concerns the service re-



System Dynamics Models, Optimization of, Figure 5
The corresponding fit to infections is poor



System Dynamics Models, Optimization of, Figure 6
Model of service delays for durable goods under warranty

quirements which can arise following the sale of a durable good. These items are typically sold with a 12-month warranty and during this time the vendor is obliged to offer service if a customer calls for it. In this particular case the vendor is not being responsive in terms of staffing the service section. The result is that as sales grow the increasing number of service requests is putting pressure on the service personnel. The delay in responding to service calls also increases and the effect of this is that future sales are depressed because of the vendor's acquired reputation for poor service response. The basic behavior mode is overshoot and collapse.

In the model depicted in Fig. 6, the growth process is achieved by a RAMP function which causes sales of the good to increase linearly by 20 units per month from a base of 500 units per month.

The payoff function is restricted to the variable *Sales*. However, this need not be the case. Where a number of variables might be options in a payoff function, it is possible to assign weights to each such that the sum of the weights is 1.0 (or 100). The optimization process will then proceed with the software accumulating a weighted payoff which it will attempt to maximize. Weights are positive when more is better and negative when less is better.

Policy Experiment No. 1

Here it is decided to try to improve the productivity of the service staff. Currently they manage, on average, to respond to 120 calls per operative per month. It may be an option to improve their productivity by, say, providing them with hand-held devices which direct each operative from one call to the next – calls which may have arisen since setting out from their base. In this way their call routing is improved.

The optimization parameter is *Prod Serv Staff*, and we select an upper limit for the search range of 240 calls per person per month. The chosen performance variable is *Sales*, since we wish to maximize this – or at least not have it overly depressed by poor response times. The results are shown in Table 5. We see that the payoff is increased and that the optimum productivity is a modest increase of 2.6 requests per month, on average. This should be easily achievable and perhaps without expenditure on high-tech devices. The graphical output for sales is shown in Fig. 7.

For comparison, the effect of increasing the productivity to as high as 150 calls per month, on average, is also shown. This would represent an increase of 25% and would be much more difficult to accomplish. Here the benefit of optimization is highlighted. A modest increase in productivity returns a visibly improved sales performance (although the basic behavior mode is unchanged), whilst a much greater productivity increase offers little extra benefit for the effort and cost involved in improving productivity.

System Dynamics Models, Optimization of, Table 5
Optimization results for the productivity of the service staff

Initial point of search
Prod Serv Staff = 120
Simulations = 1
Pass = 0
Payoff = 27 743.5

Maximum payoff found at:
*Prod Serv Staff = 122.647
Simulations = 27
Pass = 3
Payoff = 29 915

Policy Experiment No. 2

Another approach to policy optimization involves the use of a zero-one parameter which has the effect of either including or excluding an influence on policy. Suppose it was thought that the quantity of product units in warranty should exert an influence on the numbers of service personnel hired (or fired). The equation for the desired number of service staff (*Des Serv Staff*) can be expressed as:

Des Serv Staff = "Av #Serv Req Satis"/Prod Serv Staff ∗ trigger + ("Av #Serv Req Satis"/Prod Serv Staff)∗(Units Warr/initial units in warranty)∗(1-trigger). (Units: Persons).

The *trigger* variable is initially set to 1.0 and so the more sophisticated policy is not active. The optimization run results are shown in Table 6. Clearly there is benefit from including the more sophisticated policy which takes into account the current numbers of product units in warranty.

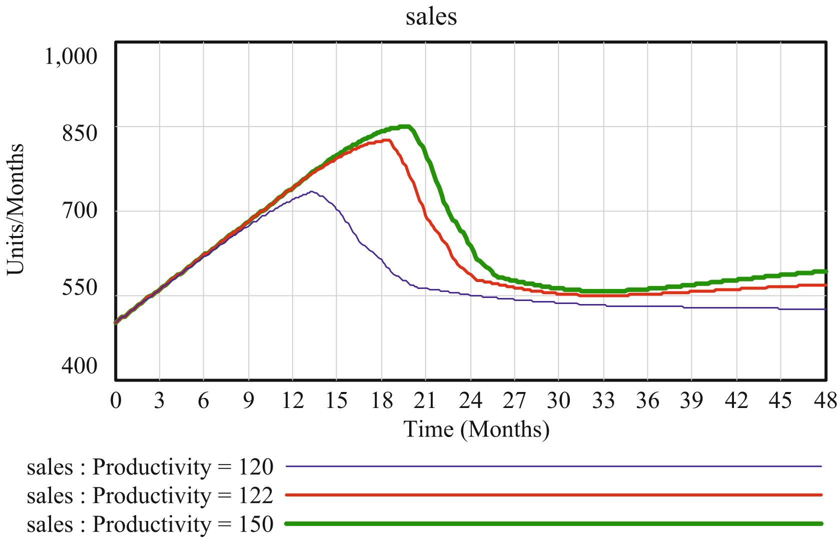
The graphical output is unequivocal (Fig. 8). Sales are continuously increasing when the recruitment policy for service personnel takes into account the number of product units in warranty. The depressive effect on sales of poor service performance is non-existent.

Whilst this might seem an obvious policy, it is surprising how easily the naïve alternative might be accepted without question. The number of calls a typical operative can manage each month is well known along with the (his-

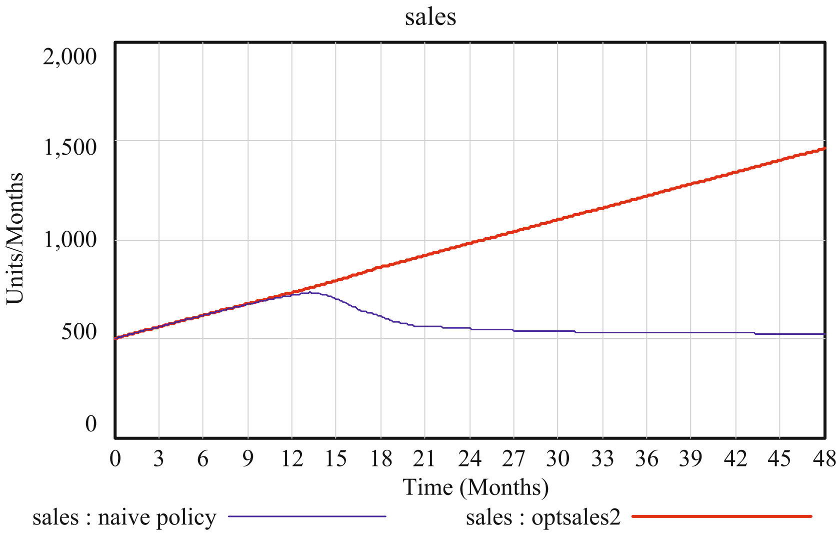
System Dynamics Models, Optimization of, Table 6
Optimization results from selection of policy drivers

Initial point of search
trigger = 1
Simulations = 1
Pass = 0
Payoff = 27 743.5

Maximum payoff found at:
∗trigger = 0
Simulations = 13
Pass = 3
Payoff = 47 090.9



System Dynamics Models, Optimization of, Figure 7
Plots of sales achieved for differing productivities



System Dynamics Models, Optimization of, Figure 8
Comparison of sales from two different policy drivers

torical) number of service requests satisfied. Hence, the desired number of staff is more or less fixed. This comes undone when there is a growth in the number of products sold. In this different environment such a simplistic policy can, as shown, lead to overshoot and collapse. Notice needs to be taken of the changing number of product units in warranty in order that a more effective system performance is achieved.

The above experiments are illustrative only, and there is no intention of over-working a simple teaching model in order to uncover an ideal policy. In the case of policy optimization a wide range of possible alternatives exists. Indeed, a process of learning naturally arises through carrying out repeated optimization experiments with the model [1].

Examples of SD Optimization Reported in the Literature

Amongst the earliest work in this area the writings of Keloharju are worthy of mention. He contributed a number of papers on the topic in the pages of *Dynamica*. See, for example, Keloharju [9]. His work brought the concept to prominence but he did not employ the method on anything other than problems described in text books or postulated by himself. For instance, an application of optimization to the project model contained in Richardson and Pugh's [13] text is contained in Keloharju and Wolstenholme [11]. A statement of the method together with some textbook examples is also available [10]. Additionally, an overview of the methods and their deployment on textbook examples has been contributed by the current author [3]. Finally, there is an example of optimization applied to defence analysis. Again though it is a standard defence model – the armored advance model – rather than any real-world study [14].

Retaining the emphasis on textbook problems for the moment, Duggan [6] employs Coyle's model [1] of the *Domestic Manufacturing Company* to illustrate the methods of multi-objective optimization – an advance over standard SD optimization with its single objective function. The concept of multiple objectives arises from multi-criteria decision-making where a situation can be judged on more than one performance metric. While a multi-objective payoff function can be formulated using a set of weights, it is argued that the selection of the weights is very individual-specific. The multi-objective approach – underpinned by the methods of genetic algorithms – rests upon determining a Pareto-optimal situation, defined as one where no improvement is possible without making some other aspect worse. In other words the method strives for an optimal solution which is not dominated by any other solution. The author demonstrates the approach combining two objectives in the model: one for the differences between desired stock and actual and another between desired backlog and actual.

In terms of applications to real-world problems, the current author has also used the methods of optimization in research conducted in connection with modeling the epidemiology of HIV/AIDS. Fitting a model of AIDS spread to data was carried out for a number of European countries [2,4]. The optimized parameters furnished support for some of the features of AIDS epidemiology which, at the time, were being uncovered by other branches of science. For example, the optimized output revealed that a U-shaped profile of infectiousness in a host was necessary in order to achieve a best fit to data on new AIDS cases. This infectiousness profile was also evidenced by virologists who had analyzed patients' blood and other secretions on a longitudinal basis.

Within this strand of research, a much more complex optimization was performed using American data on transfusion-associated AIDS cases [5]. The purpose here was to estimate the parameters for a number of plausible statistical HIV incubation distributions. Given the nature of the data, the point of infection could be quite accurately determined, but two difficulties were evident: the data were right-censored and the number receiving infected transfusions in each quarter was unknown. However, the SD optimization could estimate this number as part of the process, in addition to estimating parameters of the incubation distribution. The best fit was found to be a three stage distribution similar to the gamma and one which accorded with the high-low-high U-shaped infectiousness profile which was receiving support from a number of sources.

In the marketing domain Graham and Ariza [8] carried out an optimization on a system dynamics model which was designed to shed light on the allocations to make from a marketing budget in a high-tech client firm. Assuming the budget was fixed, the task was to optimize the allocations across more than 90 'buckets' – combinations of product lines, marketing channels and types of marketing. However, these were not discrete: advertising on one product line might have crossover effects on another and the impacts could propagate over a period of time. One major conclusion for this firm was that the advertising allocation should be increased markedly. In general intuitive allocations were shown to fall short of the ideal: they were directionally correct but magnitudes fell short often by factors of three or four.

Future Directions

A primary aim must be to see more published work which describes optimization studies carried out on real-world SD applications. There may be frequent use of optimization in consulting assignments but such activities are rarely published. The references herein suggest that, thus far, outside of unpublished work, the number may be three at most. Whilst software requirements may have inhibited use of SD optimization in the past, there are now no computational barriers to its use and it is to be hoped that in future this quite powerful analytical tool in SD will feature in more application studies.

An advance in the methodology itself has been developed by Duggan [7] and this is a promising pointer for the future. Based on genetic algorithms, it is best suited to the class of SD problems that are agent-based and this highlights a slight limitation. Traditional optimization takes the policy equations as given and explores the parameter space to determine an optimal policy. Instead he has offered an approach which searches over both parameter space and policy (strategy). Theoretically there is no limit to the number of strategies which can be evaluated in this approach, although the user has to define a set in advance of the runs. Under a conventional optimization approach a limited tilt at this is possible using the zero-one parameter method suggested above, although this would restrict the enumerated strategies to two only. Duggan demonstrates the new approach using a classic SD problem: the four agent beer-game. We await its use in a real-world application.

Bibliography

1. Coyle RG (1996) System dynamics modelling: a practical approach. Chapman & Hall, London
2. Dangerfield BC, Roberts CA (1994) Fitting a model of the spread of AIDS to data from five European countries. In: O.R. Work in HIV/AIDS, 2nd edn. Operational Research Society, Birmingham, pp 7–13
3. Dangerfield BC, Roberts CA (1996) An overview of strategy and tactics in system dynamics optimisation. J Oper Res Soc 47(3):405–423
4. Dangerfield BC, Roberts CA (1996) Relating a transmission model of AIDS spread to data: some international comparisons. In: Isham V, Medley G (eds) Models for infectious human diseases: Their structure and relation to data. Cambridge University Press, Cambridge, pp 473–476
5. Dangerfield BC, Roberts CA (1999) Optimisation as a statistical estimation tool: an example in estimating the AIDS treatment-free incubation period distribution. Syst Dyn Rev 15(3):273–291
6. Duggan J (2005) Using multiple objective optimisation to generate policy insights for system dynamics models. In: Proceedings of the international system dynamics conference, Boston. System Dynamics Society. (CD-ROM)
7. Duggan J (2008) Equation-based policy optimisation for agent-oriented system dynamics models. Syst Dyn Rev 24(1):97–118
8. Graham AK, Ariza CA (2003) Dynamic, hard and strategic questions: using optimisation to answer a marketing resource allocation question. Syst Dyn Rev 19(1)27–46
9. Keloharju R (1977) Multi-objective decision models in system dynamics. Dynamica 3(1)3–13 and 3(2)45–55
10. Keloharju R, Wolstenholme EF (1988) The basic concepts of system dynamics optimisation. Syst Pract 1:65–86
11. Keloharju R, Wolstenholme EF (1989) A case study in system dynamics optimisation. J Oper Res Soc 40(3):221–230
12. Meadows DM, Robinson JM (1985) The electronic oracle. Wiley, Chichester (Now available from the System Dynamics Society, Albany NY)
13. Richardson GP, Pugh AL (1981) An introduction to system dynamics modelling with DYNAMO. MIT Press, Cambridge (Now available from Pegasus Communications, Waltham, MA)
14. Wolstenholme EF, Al-Alusi AS (1987) System dynamics and heuristic optimisation in defence analysis, Syst Dyn Rev 3(2):102–115

System Dynamics and Organizational Learning

KAMBIZ MAANI
Chair in Systems Thinking and Practice,
The University of Queensland, Brisbane, Australia

Article Outline

Glossary

Stock In system dynamics stock is a concept representing accumulation and the state of a variable, such as, assets, inventory, capacity, reputation, morale etc. Stock can be measured at any point of time. In mathematical terms, stock is the sum over time (integral) of one or more flows.

Flow Flow or rate represents change or movement in a stock such as, buying assets, building inventories, adding capacity, losing reputation or morale, etc. Flow is measured as "per unit of time" like hiring rate (employees hired per year, production rate (units made per day), or rainfall (inches of rain per month).

Causal loops Causal loops (model) are visual maps that connect a group of variables with known or hypothesized cause and effect relationships. A causal loop can be open or closed. Causal loops can be used for complex problem solving/decision making, consensus building, conflict resolution, priority setting and group learning.

Feedback In a cause and effect chain (system), feedback is a signal from the effect/s to cause/s as to its/their influence on downstream effect/s. Feedback can be information, decision or action. For example, if X causes or changes Y, Y in turn could influence or change X directly or through other intervening variables. This creates a *closed* "causal loop" with either a positive or amplifying feedback (Reinforcing – R) or a negative feedback with damping, counteracting or (Balancing – B) effect.

Delay Cause and effect relationships are often not close in time or space. The lapse time between a cause and its effect is called a systems delay or simply delay. Because some delays in physical, natural and social systems are rather long they mask the underlying or earlier causes when effects become evident. This provides confusion and unintended consequences, especially in social systems, such as economics, education, immigration, judicial systems, etc.

Reference Mode Reference mode is the actual/observed pattern of a key variable of interest to decision makers or policy analysts. It represents the actual behavior of a variable over time which is used to compare with the simulated pattern of the same variable generated by a simulation model to validate the accuracy of the model.

Simulation A computer tool and methodology for modeling complex situations and challenging problems where mathematical tools fail to operate.

Microworld Microworlds are simulation models of real systems such as a firm, a hospital, a market, or a production system. They provide a "virtual" world where decision makers can test and experiment their policies and strategies in a laboratory environment before implementation. Microworlds are constructed using system dynamic software with user friendly interfaces.

Leverage Leverage refers to decisions and actions for change and intervention which have the highest likelihood of lasting and sustainable outcomes. Leverage decisions are best reached by open discussion after the group develops a deep understanding of system dynamics through a causal loop or stock & flow modeling process.

Systems thinking Systems thinking is a paradigm for viewing reality based on the primacy of the whole and relationships. It is one of the key capabilities (disciplines) for organizational learning [30]. Systems Thinking consists of a series of conceptual and modeling tools such as behavior over time, causal loop diagrams and systems archetypes. These tools reveal cause and effect dynamics over time and assist understanding of complex, non-linear, and counter-intuitive behaviors in all systems – physical, natural and social.

Definition of the Subject

System dynamics (SD) is "a methodology for studying and managing complex feedback systems… While the word system has been applied to all sorts of situations, feedback is the differentiating descriptor here. Feedback refers to the situation of *X* affecting *Y* and *Y* in turn affecting *X* perhaps through a chain of causes and effects… Only the study of the whole system as a feedback system will lead to correct results." [36]

Sterman ([35], p 4) defines System Dynamics as "a method to enhance learning in complex systems". "System dynamics is fundamentally interdisciplinary… It is grounded in the theory of nonlinear dynamics and feedback control developed in mathematics, physics, and engineering. Because we apply these tools to the behavior of human as well as physical and technical systems, system dynamics draws on cognitive and social psychology, economics, and other social sciences."

Wolstenholme's [40] offers the following description for system dynamics and its scope:

> A rigorous way to help thinking, visualizing, sharing, and communication of the future evolution of complex organizations and issues over time; for the purpose of solving problems and creating more robust designs, which minimize the likelihood of unpleasant surprises and unintended consequences; by creating operational maps and simulation models which externalize mental models and capture the interrelationships of physical and behavioral processes, organizational boundaries, policies, information feedback and time delays; and by using these architectures to test the holistic outcomes of alternative plans and ideas; within a framework which respects and fosters the needs and values of awareness, openness, responsibility and equality of individuals and teams.

Organizational Learning

Organizational learning is the ability of organizations to enhance their collective capacity to learn and to act, harmoniously. According to Senge [30] "Real learning gets to the heart of what it means to be human. Through learning we re-create ourselves. Through learning we become able to do something we never were able to do. Through learning we re-perceive the world and our relationship to it. Through learning we extend our capacity to create, to be part of the generative process of life. There is within each of us a deep hunger for this type of learning." Organizational learning extends this learning to the organization and its members.

Introduction

History of System Dynamics

(This section is due to US Department of Energy website) "System dynamics was created during the mid-1950s by Professor Jay W. Forrester of the Massachusetts Institute of Technology. Forrester arrived at MIT in 1939 for graduate study in electrical engineering. His first research assistantship put him under the tutelage of Professor Gordon

Brown, the founder of MIT's Servomechanism Laboratory. Members of the MIT Servomechanism Laboratory, at the time, conducted pioneering research in feedback control mechanisms for military equipment. Forrester's work for the Laboratory included traveling to the Pacific Theatre during World War II to repair a hydraulically controlled radar system installed aboard the aircraft carrier Lexington. The Lexington was torpedoed while Forrester was on board, but not sunk.

At the end of World War II, Jay Forrester turned his attention to the creation of an aircraft flight simulator for the US Navy. The design of the simulator was cast around the idea, untested at the time, of a digital computer. As the brainstorming surrounding the digital aircraft simulator proceeded, however, it became apparent that a better application of the emerging technology was the testing of computerized combat information systems. In 1947, the MIT Digital Computer Laboratory was founded and placed under the direction of Jay Forrester. The Laboratory's first task was the creation of WHIRLWIND I, MIT's first general-purpose digital computer, and an environment for testing whether digital computers could be effectively used for the control of combat information systems. As part of the WHIRLWIND I project, Forrester invented and patented coincident-current random-access magnetic computer memory. This became the industry standard for computer memory for approximately twenty years. The WHIRLWIND I project also motivated Forrester to create the technology that first facilitated the practical digital control of machine tools.

After the WHIRLWIND I project, Forrester agreed to lead a division of MIT's Lincoln Laboratory in its efforts to create computers for the North American SAGE (Semi-Automatic Ground Environment) air defense system. The computers created by Forrester's team during the SAGE project were installed in the late 1950s, remained in service for approximately twenty-five years, and had a remarkable "up time" of 99.8%.

Forrester's seminal book Industrial Dynamics [11] "is still a significant statement of philosophy and methodology in the field. Since its publication, the span of applications has grown extensively and now encompasses work in

- corporate planning and policy design
- public management and policy
- biological and medical modeling
- energy and the environment
- theory development in the natural and social sciences
- dynamic decision making
- complex nonlinear dynamics" [36]

Systems Thinking and Modeling Methodology

System Dynamics is one of the five phases of systems thinking and modeling intervention methodology [6,21]. These distinct but related phases are as follows:

1. Problem structuring;
2. Causal loop modeling;
3. System dynamics modeling;
4. Scenario planning and modeling;
5. Implementation and organizational learning (learning lab).

These phases follow a process, each involving a number of steps, as outlined in Table 1. This process does not require all phases to be undertaken, nor does each phase require all the steps listed. Which phases and steps are included in a particular project or intervention depends on the issues or problems that have generated the systems enquiry and the degree of effort that the organization is prepared to commit to.

System Dynamics Modeling

This phase follows the causal modeling phase. Although it is possible to go into this phase directly after problem structuring, performing the causal modeling phase first will enhance the conceptual rigor and learning power of the systems approach. The completeness and wider insights of systems thinking is generally absent from other simulation modeling approaches, where causal modeling does not play a part. The following steps are generally followed in the system dynamics modeling phase.

1. Develop a high-level map or systems diagram showing the main sectors of a potential simulation model, or a 'rich picture' of the main variables and issues involved in the system of interest.
2. Define variable types (e. g. stocks, flows, converters, etc.) and construct stock flow diagrams for different sectors of the model.
3. Collect detailed, relevant data including media reports, historical and statistical records, policy documents, previous studies, and stakeholder interviews.
4. Construct a computer simulation model based on the causal loop diagrams or stock-flow diagrams. Identify the initial values for the stocks (levels), parameter values for the relationships, and the structural relationships between the variables using constants, graphical relationships and mathematical functions where appropriate. This stage involves using specialized computer packages like STELLA, *ithink*, VENSIM, POWERSIM, DYSMAP, COSMIC and Consideo.

System Dynamics and Organizational Learning, Table 1
The five phase process of systems thinking and modeling (Source: [6])

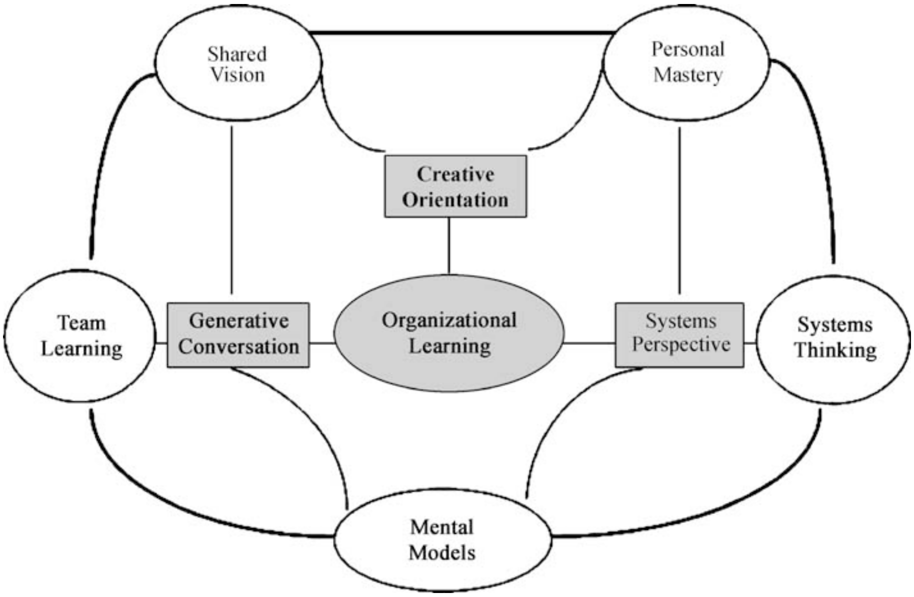
Phases	Steps	
1	Problem structuring	Identify problems or issues of concern to management, Collect preliminary information and data
2	Causal loop modeling	Identify main variables, Prepare behavior over time graphs (reference mode), Develop causal loop diagram (influence diagram), Analyze loop behavior over time and identify loop types, Identify system archetypes, Identify key leverage points, Develop intervention strategies
3	**System dynamic modeling**	Develop a systems map or rich picture, Define variable types and construct stock-flow diagrams, Collect detailed information and data, Develop a simulation model, Simulate steady -state/stability conditions, Reproduce reference mode behavior (base case), Validate the model, Perform sensitivity analysis, Design and analyze policies, Develop and test strategies
4	Scenario planning and modeling	Plan general scope of scenarios, Identify key drivers of change and keynote uncertainties, Construct forced and learning scenarios, Simulate scenarios with the model, Evaluate robustness of the policies and strategies
5	Implementation and organizational learning	Prepare a report and presentation to management team, Communicate results and insights of proposed intervention to stakeholders, Develop a microworld and learning lab based on the simulation model, Use learning lab to examine mental models and facilitate

5. Simulate the model over time. Select the initial value for the beginning of the simulation run, specify the unit of time for the simulation (e. g. hour, day, week, month, year, etc.). Select the simulation interval (DT) (e. g. 0.25, 0.5, 1.0) and the time horizon for the simulation run (i. e. the length of the simulation). Simulate model stability by generating steady state conditions.
6. Produce graphical and tabular output for the base case of the model. This can be produced using any of the computer packages mentioned above. Compare model behavior with historical trends or hypothesized reference modes (behavior over time charts).
7. Verify model equations, parameters and boundaries, and validate the model's behavior over time. Carefully inspect the graphical and tabular output generated by the model.
8. Perform sensitivity tests to gauge the sensitivity of model parameters and initial values. Identify areas of greatest improvement (key leverage points) in the system.
9. Design and test policies with the model to address the issues of concern to management and to look for system improvement.
10. Develop and test strategies (i. e. combinations of functional policies, for example operations, marketing, finance, human resources, etc.).

Organizational Learning

(This section is adapted from [21])
Peter Senge, who popularized the concept through his seminal book: The Fifth Discipline [30], describes a learning organization as one 'which is continually expanding its ability to create its future'. He identifies five core capabilities (disciplines) of the learning organization that are derived from three "higher orientations": creative orientation; generative conversation; and systems perspective. "The reality each of us sees and understands depend on what we believe is there. By learning the principles of the five disciplines, teams begin to understand how they can think and inquire that reality, so that they can collaborate in discussions and in working together create the results that matter [to them]."

As Fig. 1 shows the learning organization capabilities are dynamically interrelated, and collectively they lead to organizational learning.



System Dynamics and Organizational Learning, Figure 1
The core capabilities of a learning organization (Source: [19])

Senge maintains that *Creative orientation* is the source of a genuine desire to excel. It is the source of an intrinsic motivation and drive to achieve. It relinquishes personal gains in favor of the common good. *Generative conversation* refers to a deep and meaningful dialog to create unity of thought and action. *Systems perspective* is the ability to see things holistically by understanding the interconnectedness of the parts. The foregoing elements give rise to the five core capabilities of learning organizations, namely: personal mastery; shared vision; mental models; team learning and dialog; and systems thinking. These five disciplines are described below. Figure 1 below shows the core capabilities and their relationships.

Personal Mastery

Senge [30] describes that personal mastery is the cornerstone and 'spiritual' foundation of the learning organization. It is born out of a creative orientation and systemic perspective. Personal mastery instils a genuine desire to do well and to serve a noble purpose. People exhibiting high levels of personal mastery focus "on the desired result itself, not the process or the means they assume necessary to achieve that result" [30]. These people can "successfully focus on their ultimate intrinsic desires, not on secondary goals. This is a cornerstone of Personal Mastery". Personal mastery also requires a commitment to truth, which means to continually challenge "theories of why things are the way they are". Without committing to the truth, people all too quickly revert to old communication routines which can distort reality and prevent them from knowing where they really stand.

Shared Vision

It is commonly assumed that in contemporary organizations senior management can develop a vision which employees will follow with genuine commitment. This is a fallacy. Simply promoting a 'vision statement' could result in a sense of apathy, complacency and resentment. Instead, there needs to be a genuine endeavor to understand what people will commit to. The overriding vision of the group must build on the personal visions of its members. Shared vision should align diverse views and feelings into a unified focus.

This is emphasized by Arie de Geus [9] when he describes what makes a truly extraordinary organization. "The feeling of belonging to an organization and identifying with its achievements is often dismissed as soft. But case histories repeatedly show that a sense of community is essential for long term survival". For example, when Apple Corporation challenged IBM, it was in its 'adolescent' years, characterized by creativity, confidence and even defiance. This is similar to the spirit in Team New Zealand when it competed against the bigger-budget syndicates! Within these organizations there is a real passion for the outcome; a common vision for success [19].

Creating a shared vision is the most fundamental job of a leader [26]. By creating a vision, the leader provides a vehicle for people to develop commitment, a common goal around which people can rally, and a way for people to *feel* successful. The leader must appeal to people's emotions if they are to be energized towards achieving the goal. Emotional acceptance of, and belief in, a vision is far more powerful in energizing team members than is intellectual recognition that the vision is simply a 'good idea'. One of the most powerful ways of communicating a vision is through a leader's personal example and actions, demonstrating behavior that symbolizes and furthers that vision.

Mental Model and Leadership

Mental models reflect beliefs, assumptions and feelings that shape one's world views and actions. They are formed through family, education, professional and social learning based, on the most part, on cultural and social norms. Mental models, however, can be altered and aligned.

Organizations are often constrained by deep-seated belief systems, resulting in preconceived ideas on how things ought to perform. Goodstein and Burke ([14] p. 10), pioneers in the field of social psychology of organizations, observed that 'the first step in any change process is to unfreeze the present patterns of behavior as a way of managing resistance to change'. The leader has a pivotal role in dismantling negative mental models and shaping new ones.

In order to get people to engage in open discussions of issues that affect the organization, a leader must appeal to their emotions and must get beyond the superficial level of communication. In the 1970s Shell Oil undertook major changes in its leadership approach and communications style. According to a manager at Shell, "When I tried to talk personally about an issue rather than say 'here's the answer', it was powerful. It caused me to engage in dialog with others that resulted in mutual learning on all sides" ([7] p. 71).

The leader is a 'designer', and part of that role is designing the governing ideas of purpose and core values by which people will live [30,31]. In this role, the leader must propose and model the manner in which the group has to operate internally. This provides ample opportunities for leaders to examine their deeply held assumptions about the task, the means to accomplish it, the uniqueness of the people and the kinds of relationship that should be fostered among the people. Only after people have observed and *experienced* the organizational values in practice would these values become the basis for prolonged group behavior. These values should be manifested first and should be most visible in the leader's own behavior.

Leadership, especially in knowledge-based organizations, must be distributed and shared to a far greater extent than it was in the past. For example, in the Chicago Bulls basketball team, Michael Jordan changed his role: it became not only that of an individually brilliant player but *also* that of a leader whose job it was to raise the level of play of other team members. After this transition, the Bulls began their record run of championship seasons [7].

Team Learning and Dialog

The word 'dialog' comes from the Greek words *dia* and *logos*. It implies that when people engage in dialog, the meaning *moves through* them – Thus, it enables them to 'see through words' [16]. Dialog is an essential requirement for organizational learning. It results from generative conversation, shared vision, and transparent mental models. Dialog creates a deep sense of listening and suspending one's own views. Feedback is an integral aspect of dialog.

Communication routines in organizations are generally anti-learning and promote mediocrity. They include 'defensive routines' [2] – statement that can stifle dialog and innovative thinking. Exposing and unlearning such routines, and understanding the powerful detrimental impact they have on learning, are serious challenges many organizations face if they are to create effective learning environments.

Many leaders are charismatic and are highly eloquent when it comes to presenting their ideas; that's often why they get to the top of the organization. However, many appear to lack the ability to extract the very best from employees in a non-threatening manner. Without this ability, leaders may miss many good ideas, or might act on many bad ones.

In a group context, encouragement from the leader and mutual encouragement among group members is essential. Furthermore, personal differences must be put aside in order for effective dialog to ensue.

How Organizations Learn

(This section is edited from Wikipedia: http://en.wikipedia.org/wiki/Organizational_learning)

"Argyris and Schon were the first to propose concepts and models that facilitate organizational learning, the following literatures have followed in the tradition of their work:

- March and Olsen [23] attempt to link up individual and organizational learning. In their model, individual beliefs lead to individual action, which in turn may lead to an organizational action and a response from the environment which may induce improved individual beliefs and the cycle then repeats over and over. Learning occurs as better beliefs produce better actions.

- Argyris and Schon [3] distinguish between single-loop and double-loop learning, related to Gregory Bateson's concepts of first and second order learning. In single-loop learning, individuals, groups, or organizations modify their actions according to the difference between expected and obtained outcomes. In double-loop learning, the entities (individuals, groups or organization) question the values, assumptions and policies that led to the actions in the first place; if they are able to view and modify those, then second-order or double-loop learning has taken place. Double loop learning is the learning about single-loop learning.
- Kim [17], as well, in an article titled "The link between individual and organizational learning", integrates Argyris, March and Olsen and another model by Kofman into a single comprehensive model; Further, he analyzes all the possible breakdowns in the information flows in the model, leading to failures in organizational learning; For instance, what happens if an individual action is rejected by the organization for political or other reasons and therefore no organizational action takes place?
- Nonaka and Takeuchi [27] developed a four stage spiral model of organizational learning. They started by differentiating Polanyi's concept of "tacit knowledge" from "explicit knowledge" and describe a process of alternating between the two. Tacit knowledge is personal, context specific, subjective knowledge, whereas explicit knowledge is codified, systematic, formal, and easy to communicate. The tacit knowledge of key personnel within the organization can be made explicit, codified in manuals, and incorporated into new products and processes. This process they called "externalization". The reverse process (from explicit to implicit) they call "internalization" because it involves employees internalizing an organization's formal rules, procedures, and other forms of explicit knowledge. They also use the term "socialization" to denote the sharing of tacit knowledge, and the term "combination" to denote the dissemination of codified knowledge. According to this model, knowledge creation and organizational learning take a path of socialization, externalization, combination, internalization, socialization, externalization, combination… etc. in an infinite spiral.
- Flood [10] discusses the concept of organizational learning from Peter Senge and the origins of the theory from Argyris and Schon. The author aims to "re-think" Senge's *The Fifth Discipline* through systems theory. The author develops the concepts by integrating them with key theorists such as Bertalanffy, Churchman, Beer, Checkland and Ackoff. Conceptualizing organizational learning in terms of structure, process, meaning, ideology and knowledge, the author provides insights into Senge within the context of the philosophy of science and the way in which systems theorists were influenced by twentieth-century advances from the classical assumptions of science.
- Nick Bontis et al. [4] empirically tested a model of organizational learning that encompassed both stocks and flows of knowledge across three levels of analysis: individual, team and organization. Results showed a negative and statistically significant relationship between the misalignment of stocks and flows and organizational performance.
- Imants [15] provides theory development for organizational learning in schools within the context of teachers' professional communities as learning communities, which is compared and contrasted to teaching communities of practice. Detailed with an analysis of the paradoxes for organizational learning in schools, two mechanisms for professional development and organizational learning, (1) steering information about teaching and learning and (2) encouraging interaction among teachers and workers, are defined as critical for effective organizational learning.
- Common [8] discusses the concept of organizational learning in a political environment to improve public policy-making. The author details the initial uncontroversial reception of organizational learning in the public sector and the development of the concept with the learning organization. Definitional problems in applying the concept to public policy are addressed, noting research in UK local government that concludes on the obstacles for organizational learning in the public sector: (1) overemphasis of the individual, (2) resistance to change and politics, (3) social learning is self-limiting, i. e. individualism, and (4) political "blame culture". The concepts of *policy learning* and *policy transfer* are then defined with detail on the conditions for realizing organizational learning in the public sector."

Modeling for Organizational Learning

In general, the *process* of model building can be an effective conduit for collective learning. System Dynamics modeling, in particular, can be used to enhance organizational learning [35] through rapid feedback and experimentation and its facility to test assumptions and mental models. As we have discussed, dealing effectively with mental models is one of the core competencies for organizational learning

Ackoff [1] likens complex problems to "messes". "Messy problems are defined as situations in which there are large differences of opinion about the problem or even on the question of whether there is a problem. Messy situations make it difficult for a management team to reach agreement. System Dynamics modeling with groups known as Group Model Building (GMB) is a powerful tool for dealing with these. SD and GMB are especially effective in dealing with semi-structured and ill-structured decision situations."

GMB offers an opportunity to align and share piecemeal mental models and create the possibility of assimilating and integrating partial mental models into a holistic system description [38,39]. GMB and SD can help uncover 'illusions' that may occur due to the fact that the definition of a problem may be a socially constructed phenomenon that has not been put to test ([18] p. 84).

Learning Laboratory

(This section is adapted from [21], Chapter 6)
Learning laboratory is a setting as well as a process in which a group can learn together. The purpose of the learning lab is to enable managers to test their long held assumptions and to experiment and 'see' the consequences of their actions, policies and strategies. This often results in finding inconsistencies and the discovery of *unintended* consequences of actions and decisions, *before* they are implemented. System Dynamics models known as Microworlds or Management Flight Simulators (MFS) are the 'engine' behind the learning lab. "Just as an airline uses flight simulators to help pilots learn, system dynamics is, partly, a method for developing management flight simulators, often computer simulation models, to help us learn about dynamic complexity, understand the sources of policy resistance, and design more effective policies." ([35], p. 4)

A learning lab is distinct from so-called management games. In management games, the players are required to compete – design the 'best' strategy and 'beat' other players or teams. The competitive nature of management games often encourages aggressive and individualistic behavior with scant regard for group learning and gaining deep insights. The learning lab, in contrast, aims to enhance *learning*: To test individual and group mental models and to provide deeper understanding and insights into why systems behave the way they do. This will help the participants to test their theories and discover inconsistencies and 'blind spots' in policies and strategies *before* they are implemented.

A significant benefit of the learning lab stems from the process in which participants examine, reveal and test their mental models and those of their organization. The learning lab can also help participants

- To align strategic thinking with operational decisions;
- To connect short-term and long-term measures;
- To facilitate integration within and outside the organization;
- To undertake experimentation and learning;
- To balance competition with collaboration.

Managerial Practice Field

Team and teamwork are parts of the lexicons of numerous organizations today. Company after company has reorganized work around a variety of team concepts. From factories to hospitals, *titles* like 'manager' and 'supervisor' have been replaced by *roles* such as 'facilitator' and 'team leader'. Despite this level of attention to team and teamwork the expected benefits have been marginal at best.

But when we examine real teams, such as sporting teams, orchestras or ballet companies more closely, they all share one key characteristic. That is they *practice* a lot more than they 'perform'. Practice involves allowing time and space to experiment with new ways, try different approaches and most importantly, make mistakes without the fear of failure. In fact, making mistakes is indispensable to learning. One cannot learn from doing things right all the time! Yet a great deal of organizational energy and attention is devoted to the prevention and masking of mistakes.

But, what is the *practice field* for management teams? The fact is that the practice field is, by and large, absent from the managerial world. In other words, there is no time and no space for management to 'practice' in the true sense of the word – to experiment, make mistakes and learn together. In this era of restructuring and downsizing, lack of time is the greatest impediment to managerial and organizational learning. As a recent advertisement by IBM reads, "Innovative Thinking! We don't even have time for bad thinking". The pace in the modern work environment is so unrelenting that there is virtually no room for managers to slow down, to practice, to reflect and learn. The consequence of this lack of practice and learning space is grave, in that most organizations only achieve a small fraction of their potential – about 5%, according to Jay Forrester, the father of System Dynamics [13].

In order to fill this gap, the concept of learning laboratory has been developed to provide practice fields for managers. The learning lab allows learning to become an integral part of managerial work and helps learning to become institutionalized [17].

Aligning Mental Models Through the Learning Laboratory

Mental models are formed throughout one's life. Family, school, culture, religion, profession and social norms play important roles in this formation. Therefore, modifying one's mental model is not a small matter. The most effective way to check one's mental models is to *experience* alternative realities at first hand and see their implications with a new 'lens' [5].

There are rarely any opportunities in the course of a manager's daily work for him/her to engage in lengthy, drawn-out experimentation. Learning in a 'laboratory' setting is a viable and powerful alternative. Fortunately, advanced computers and sophisticated system dynamics software have enabled the creation of managerial learning labs where managers can experiment, test their theories and learn rapidly. Thus, learning labs can play a significant role in clarifying and changing mental models. Learning lab deals with mental models at three levels [33], as described below.

- *Mapping* mental models. This step begins at the conceptualization phase. Here, the learning lab participants articulate and clarify their assumptions, views, opinions, and biases regarding the issue at hand.
- *Challenging* mental models. The participants identify and discuss inconsistencies and contradictions in their assumptions. This step will begin at the conceptualizations phase and will continue to the experimentation phase.
- *Improving* mental models. Having conducted experimentation and testing, the participants reflect on the outcomes. This may cause them to alter, adjust, improve and harmonize their mental models.

The laboratory setting provides a neutral and 'safe' space for the participants to create a shared understanding of complex and endemic issues. The following characteristics of the learning lab provide a powerful catalyst for alignment of divergent mental models in the organization.

- The laboratory environment is neutral and non-threatening. The emphasis is on learning and theory building (what we *don't* know), not on winning or display of knowledge.
- Lack of hierarchy. Managers and staff are equal in this environment. The traditional hierarchy is minimized in the laboratory setting.
- The response time is fast. Hence, the feedback cycle is short, which leads to rapid learning.
- There is no cost or 'loss of face' attached to failure. Hence, it is safe to make mistakes. In fact, mistakes provide opportunities for learning.
- People can see the consequences of their actions first hand. No one attempts to convince or teach anyone else or force his or her preconceived views on others. People learn by themselves and through group interactions.

Implications for Management

The practice field and the learning lab concepts offer fresh and challenging implications for managers and their role. They suggest that a leader/manager should think as a *scientist*, be open to and welcome hard questions, experiment with new ideas, and be prepared to be *wrong*. This requires managers to learn systems thinking skills and use them not just for 'solving' problems but as powerful tools for communication, team building and organizational learning. This means that an effective leader should be the 'designer' of the ship and not its captain [31]. Once they have designed a new structure, strategy, policy or procedure then the managers/leaders should allow (i. e. create a practice field for) the staff to experience the new design, and experiment with it and learn for themselves – the desired outcome is shared understanding leading to alignment of thoughts and actions. This is the essence of organizational learning.

Future Directions

Agent-Based Modeling (ABM)

Agent based modeling (ABM) is an emerging modeling technology which draws its theories and techniques from complexity science [29]. While System Dynamics and Agent-Based Modeling (ABM) use different modeling philosophies and approaches, they can be used complementarily and synergistically.

System Dynamics focuses on modeling structures (i. e. relationships, policies, strategies) that underlie behavior of systems. This may be viewed as a weakness of system dynamics approach in that behavior is assumed to be solely a function of structure (model relationships defined a priori). In contrast, in ABM, organizations are modeled as a system of semi-autonomous decision-making elements – purposeful individuals called *agents*. Each agent individually assesses its situation and makes decisions based upon value hierarchies representing goals, preferences, and standards for behavior. Thus, macro-behavior is not modeled separately but *emerges* from the micro-decisions of individual agents. In other words, in agent based modeling; "emergent" behavior is expected as a result of agents' interactions. This is a key difference between the two approaches.

While system dynamics acknowledges the critical role of individual and organizational mental models (e. g., motivations, values, norms, biases, etc.) it does not explicitly model them. SD utilizes factual data or "cold knowledge" and does not take into account decision makers 'mood'. In contrast, ABM attempts to capture "warm knowledge", representing emotional and human context of decision-making.

Recent advances in video game technology allow the development of multi-agent, artificial 'society' simulators with capabilities for modeling physiology, stress and emotion in decision-making [34]. At the simplest level, an agent-based model consists of a system of agents and their relationships. This new approach enables superior understanding of the complexity in organizations and their relevant business environments. This in turn provides an opportunity for new sophistications in game-play that enhances decision-making. Experience with agent-based modeling shows that even a simple agent-based model can exhibit complex behavior patterns and provide valuable information about the dynamics of the real world system that emulates them.

Despite their differences, SD and ABM can be used in a complementary fashion. Both ABM and SD are powerful tools for transforming information into knowledge and understanding leading to individual and group learning. However, the transition from knowledge to understanding may not be immediate or transparent. This requires a deep shift in mental models through experimentation and group learning.

Systems Thinking and Sustainability

Systems Thinking has a natural affinity with sustainability modeling and management. Sustainability issues are complex; cut across several disciplines; involve multiple stakeholders and require a long term integrated approach. Thus, the systems paradigm and tools have direct and powerful applications in sustainability issues and management.

The applications of system dynamics in sustainability go back to the early 1970s with Jay Forrester's "World2 and World3 analyzes against 30 years of history", followed by "World Dynamics" and "Limits to Growth" [24] and "Beyond the Limits" [25]. "The politics of the environment has also evolved dramatically since 1970. Public awareness of the reality of the environmental challenge has risen; Ministries of Environment have become commonplace" ([28] p. 220). As an example today concern over carbon emissions has already become an international currency. As a result, sustainability has brought a fresh challenge for governments, business and industry, scientists, farmers and all the citizens of the world collectively to find systemic solutions that are mutually and globally agreeable. Systems Thinking and System dynamics can make real and valuable contributions to addressing this challenge.

Bibliography

Primary Literature

1. Ackoff RA (1999) Re-creating the corporation – A design of organizations for the 21st century. Oxford University Press, Oxford
2. Argyris C (1992) The next challenge for TQM: Overcoming organisational defences. J Qual Particip 15:26–29
3. Argyris C, Schon D (1978) Organizational learning: A theory of action perspective. Addison-Wesley, Reading
4. Bontis N, Crossan M, Hulland J (2002) Managing an organizational learning system by aligning stocks and flows. J Manag Stud 39(4):437–469
5. Brown JS (1991) Research that reinvents the corporation. Harv Bus Rev 68:102–111
6. Cavana R, Maani K (2004) A methodological framework for integrating systems thinking and system dynamics. In: System Dynamics Society Proceedings. Oxford
7. Cohen E, Tichy N (1997) How leaders develop leaders. Training and Development, May
8. Common R (2004) Organisational learning in a political environment: Improving policy-making in UK government. Policy Stud 25(1):35–49
9. De Geus A (1997) The living company. Harv Bus Rev 75(2):51–59
10. Flood RL (1999) Rethinking the fifth discipline: Learning within the unknowable. Routledge, London
11. Forrester JW (1961) Industrial dynamics. Productivity Press, Cambridge
12. Forrester JW (1971) World dynamics. Wright-Allen (Subsequently re-published by Productivity Press, and Pegasus Comunications)
13. Forrester JW (1994) Building a foundation for tomorrow's organizations. In: Systems thinking in action video collection, vol 1. Pegasus Communications, Cambridge
14. Goodstein L, Burke W (1991) Creating successful organisation change. Organ Dyn 19(4):5–17
15. Imants J (2003) Two basic mechanisms for organizational learning in schools. Europ J Teach Educ 26(3):293–311
16. Isaacs W (1993) Taking flight: Dialogue, collective thinking and organisational learning. Organ Dyn 22(2):24–39
17. Kim DH (1993) The link between individual and organizational learning. Sloan Manag Rev 35(1):37–50

18. Maani K (2002) Consensus building through systems thinking – the case of policy and planning in healthcare. Aust J Inform Syst 9(2):84–93
19. Maani K, Benton C (1999) Rapid team learning. Lessons from team New Zealand's America's cup campaign. Organ Dyn 27(4)
20. Maani K, Cavana R (2007) Systems methodology. Syst Think 18(8):2–7
21. Maani K, Cavana R (2007) Systems thinking, system dynamics – Managing change and complexity, 2nd edn. Prentice Hall, Pearson Education, Auckland
22. Maani K, Pourdehnad J, Sedehi H (2003) Integrating system dynamics and intelligent agent-based modelling – theory and case study. Euro INFORMS, Istanbul
23. March JG, Olsen JP (1975) The uncertainty of the past; Organizational ambiguous learning. Europ J Polit Res 3:147–171
24. Meadows DH, Meadows DL, Randers J, Behren W (1972) The limits to growth. Universe Press, New York
25. Meadows DH, Meadows DL, Randers J (1992) Beyond the limits. Chelsey Green, Post Mills
26. Nadler DA, Tushman ML (1990) Beyond the charismatic leader: Leadership and organisational change. Calif Manag Rev
27. Nonaka I, Takeuchi H (1995) The knowledge creating company. Oxford University Press, New York
28. Randers J (2000) From limits to growth to sustainable development or SD (sustainable development) in a SD (system dynamics) perspective. Syst Dyn Rev 16(3):213–224
29. Rothfeder J (2003) Expert voices: Icosystem's Eric Bonabeau. CIO Insights
30. Senge P (1990) The fifth discipline: The art and practice of the learning organisation. Currency
31. Senge P (1990) The leader's New Work: Building learning organisation's. Sloan Manag Rev:7–23
32. Senge P (1992) Building learning organisation's. J Qual Particip:1–8
33. Senge P, Sterman JD (1991) Systems thinking and organizational learning: Acting locally and thinking globally in the organization of the future. In: Kochan T, Useem M (eds) Transforming organizations. Oxford University Press, Oxford
34. Silverman BG et al (2002) Using human models to improve the realism of synthetic agents. Cogn Sci Q 3
35. Sterman JD (2000) Business dynamics, systems thinking and modeling for a complex world. McGraw-Hill, Irwin
36. System Dynamics Society website http://www.systemdynamics.org/
37. US Department of Energy Introduction to system dynamics, A systems approach to understanding complex policy issues, US Department of Energy. http://www.systemdynamics.org/DL-IntroSysDyn/inside.htm
38. Vennix JAM (1995) Building consensus in strategic decision-making: System Dynamics As A Support System. Group Decis Negot 4(4):335–355
39. Vennix JAM (1996) Group model-building: Facilitating team learning using system dynamics. Wiley, Chichester, chapt 5
40. Wolstenholme E (1997) System dynamics in the elevator (SD1163), e-mail communication, 24 Oct 1997 system-dynamics@world.std.com

Books and Reviews

(This section is due to M. Anjali Sastry and John D. Sterman, "An Annotated Survey of the Essential System Dynamics Literature System Dynamics Group", Sloan School of Management, MIT)

Industrial and Economic Dynamics: The Foundations

Forrester JW (1961) Industrial dynamics. Productivity Press, Cambridge (Presents dynamic analysis of a business problem through a model of a production-distribution system that shows oscillatory behavior. Policies to improve system performance are discussed, and numerous policy experiments are demonstrated. Includes full equation listing.)

Forrester JW (1968) Principles of systems. Productivity Press, Cambridge (System structure and behavior are differentiated, with examples showing how structure determines behavior. Rates and levels are described. Inventory model shows effects of delivery delay and resulting production cycles.)

Forrester JW (1975) Collected papers of Jay W. Forrester. Productivity Press, Cambridge (Includes many seminal papers, such as Industrial Dynamics: A Major Breakthrough for Decision Makers; Common Foundations Underlying Engineering and Management; A New Corporate Design; Market Growth as Influenced by Capital Investment; and Counterintuitive Behavior of Social Systems.)

Forrester JW (1989) The beginnings of system dynamics (Working Paper No. D-4165). System Dynamics Group, Sloan School of Management, MIT, Cambridge (A personal history beginning on the high plains of western Nebraska. Describes the early projects that shaped the field.)

Mass NJ (1975) Economic cycles: An analysis of underlying causes. Productivity Press, Cambridge (Shows how production scheduling and work force management policies generate the 3–5 year business cycle. Economic cycles, in turn, are caused by capital investment policies that fail to account for delays in acquiring long-lead time plant and equipment.)

Meadows DL (1970) Dynamics of commodity production cycles. Productivity Press, Cambridge (Develops a simple generic model of commodity supply and demand with explicit production capacity and delays, prices and markets. Applies the model to hogs, chicken and cattle.)

Urban and Public Policy Dynamics

Alfeld LE, Graham AK (1976) Introduction to urban dynamics. Productivity Press, Cambridge (A very readable introductory text. Uses the urban system as an example to teach general points about modeling methods, formulation and analysis.)

Forrester JW (1969) Urban dynamics. Productivity Press, Cambridge (Seminal model of urban growth and decay, controversial then and vindicated now. Chapter 6 describes general characteristics of complex systems such as compensating feedback and shifting the burden to the intervener.)

Mass NJ (ed) (1974) Readings in urban dynamics, vol I. Productivity Press, Cambridge (Extensions, modification, and responses to criticisms of the Urban Dynamics model.)

Schroeder WW, III, Sweeney RE, Alfeld LE (eds) (1975) Readings in urban dynamics, vol II. Productivity Press, Cambridge (Further extends and explores the Urban Dynamics model.)

Limits to Growth and Other Global Models

Forrester JW (1973) World Dynamics, 2nd edn. Productivity Press, Cambridge (The first global model, on which Limits to Growth was based. The extreme simplicity of the model allowed it to be presented to a wide audience.)

Meadows DL, Meadows DH (eds) (1974) Toward global equilibrium: Collected papers. Productivity Press, Cambridge (Describes and explores, through system dynamics models, policies for sustainability designed to avoid the collapse shown in the 'business as usual' WORLD3 scenarios.)

Meadows DH, Meadows DL, Randers J, Behrens WW III (1972) The limits to growth: A report for the club of Rome's project on the predicament of mankind. Universe Books, New York (Classic controversial study of the human future. Nontechnical presentation of structure, assumptions, and results of the WORLD3 model. Concluded that present policies were unsustainable; shows how alternate policies could stabilize population at a high standard of living.)

Meadows DL, Behrens WW III, Meadows DH, Naill RF, Randers J, Zahn EKO (1974) Dynamics of growth in a finite world. Productivity Press, Cambridge (Full documentation and data for the WORLD3 model used in the Limits to Growth. Describes the structure and assumptions; includes all data needed for complete replication of all runs in the popular book. Formulations described here may be useful to all system dynamics modelers.)

Meadows D, Richardson J, Bruckmann G (1982) Groping in the dark. Wiley, New York (Describes a range of global models built under different approaches and discusses the strengths, weaknesses, and implications of each. Presented in an engaging, personal style.)

Meadows DH, Meadows DL Randers J (1992) Beyond the limits: Confronting global collapse, envisioning a sustainable future. Chelsea Green, Post Mills (Follows up on Limits to Growth. Shows that many problems described in 1972 have worsened, as predicted by the model. Argues for a shift in values necessary to create a sustainable and equitable future.)

SD for Management: Firm and Market Models

Coyle RG (1977) Management System Dynamics. Wiley, New York (Text emphasizing managerial modeling, with a focus on operations and examples including discrete elements.)

Hall RI (1976) A system pathology of an organization: The rise and fall of the Old Saturday Evening Post. Adm Sci Q 21(2):185–211 (A case-study using a system dynamics model to explain how failure to understand the feedbacks among policies governing ad rates, ad and editorial pages, marketing, and pricing lead to the failure of the Post just as circulation reached an all-time high.)

Lyneis JM (1980) Corporate planning and policy design. Productivity Press, Cambridge (Begins with a simple model of inventory management in a manufacturing firm and gradually extends the model to one of the entire firm.)

Merten PP (1991) Loop-based strategic decision support systems. Strat Manag J 12:371–382 (Describes a model of a multinational firm establishing new markets in less-developed countries. Captures qualitative shifts in firm structure and organization endogenously as the firm evolves.)

Morecroft JDW (1984) Strategy support models. Strat Manag J 5(3):215–229 (Describes the use of models as participants in the ongoing dialogue among managers regarding strategy formation and evaluation. Emphasizes the processes for model development and use that enhance the utility of modeling in design of high-level corporate strategy.)

Morecroft JDW, Lane DC, Viita PS (1991) Modelling growth strategy in a biotechnology startup firm. Syst Dyn Rev 7(2):93–116 (Describes a case-study of a start-up in which system dynamics modeling helps to define a desirable growth strategy for the firm. The integrated model generated strategies that allowed different parts of the firm to choose consistent approaches.)

Roberts EB (ed) (1978) Managerial applications of system dynamics. Productivity Press, Cambridge (Extensive collection of early corporate models, including history and commentary by practitioners. Covers R&D management, production and operations, human resources, and other applications areas.)

Economic Models

Forrester JW (1989) The system dynamics national model: Macrobehavior from microstructure. In: Milling PM, Zahn EOK (eds) Computer-based management of complex systems: International System Dynamics Conference. Springer, Berlin (Provides an overview of the national modeling project in which both micro- and macro-economic factors are included. Model generates business cycles, inflation, stagflation, the economic long wave, and growth.)

Saeed K (1986) The dynamics of economic growth and political instability in the developing countries. Syst Dyn Rev 2(1):20–35 (Shows how rapid economic development can generate social and political instability through a model that links socio-political factors to economic development.)

Sterman JD (1985) A behavioral model of the economic long wave. J Econ Behav Organ 6(1):17–53 (Proposes and tests a simple model of the long wave. The intended rationality of each decision rule is tested and the long wave is explained as the unintended result of the interaction of locally rational decision processes. The model is the basis for the STRATAGEM-2 game, and can exhibit chaos.)

Sterman JD (1989) Deterministic chaos in an experimental economic system. J Econ Behav Organ 12:1–28 (Sterman's 1985 model of the long wave is converted into a management flight simulator and used as an experiment in which subjects make the capital investment decision. Simple decision rules capturing subject's policies are estimated and explain their behavior well. Simulation of these rules yields deterministic chaos for about 25% of the subjects.)

Sterman JD (1986) The economic long wave: Theory and evidence. Syst Dyn Rev 2(2):87–125 (Comprehensive overview of the theory of long waves arising from the System Dynamics National Model. Reviews the feedback structures responsible for the long wave and empirical evidence supporting the dynamic hypotheses. Discusses the role of innovation and political value change.)

Conceptualizing, Formulating and Validating Models

Barlas Y (1989) Multiple tests for validation of system dynamics type of simulation models. Europ J Operat Res 42(1):59–87 (Discusses a variety of tests to validate SD models, including structural and statistical tests.)

Barlas Y, Carpenter S (1990) Philosophical roots of model validation: Two paradigms. Syst Dyn Rev 6(2):148–166 (Contrasts the system dynamics approach to validity with the traditional, logical empiricist view of science. Finds that the relativist philosophy is consistent with SD and discusses the practical implications for modelers and their critics.)

Forrester JW (1980) Information sources for modeling the national economy. J Am Stat Assoc 75(371):555–574 (Argues that modeling the dynamics of firms, industries, or the economy requires use of multiple data sources, not just numerical data and statistical techniques. Stresses the role of the mental and descriptive data base; emphasizes the need for first-hand field study of decision making.)

Forrester JW (1985) The model versus a modeling process. Syst Dyn Rev 1(1):133–134 (The value of a model lies not in its predictive ability alone but primarily in the learning generated during the modeling process.)

Forrester JW (1987) Fourteen 'Obvious Truths'. Syst Dyn Rev 3(2):156–159 (The core of the system dynamics paradigm, as seen by the founder of the field.)

Forrester JW (1987) Nonlinearity in high-order models of social systems. Europ J Operat Res 30(2):104–109 (Nonlinearity is pervasive, unavoidable, and essential to the functioning of natural and human systems. Modeling methods must embrace nonlinearity to yield realistic and useful models. Linear and nearly-linear methods are likely to obscure understanding or lead to erroneous conclusions.)

Homer JB (1983) Partial-model testing as a validation tool for system dynamics. In: International System Dynamics Conference, pp 920–932 (How model validity can be improved through partial model testing when data for the full model are lacking.)

Legasto AA Jr, Forrester JW, Lyneis JM (eds) (1980) System dynamics. In: TIMS studies in the management sciences, vol 14. North-Holland, Amsterdam (Collection of papers focused on methodology. Includes Forrester and Senge on Tests for Building Confidence in System Dynamics Models and Gardiner & Ford's discussion on Which Policy Run is Best, and Who Says So?)

Mass N (1991) Diagnosing surprise model behavior: A tool for evolving behavioral and policy insights. Syst Dyn Rev 7(1):68–86 (Guidelines for learning from surprise model behavior with tests to resolve anomalous behavior.)

Morecroft JDW (1982) A critical review of diagramming tools for conceptualizing feedback system models. Dynamica 8(1):20–29 (Critiques causal-loop diagrams and proposes subsystem and policy structure diagrams as superior tools for representing the structure of decisions in feedback models.)

Randers J (ed) (1980) Elements of the system dynamics method. Productivity Press, Cambridge (Includes Mass on Stock and Flow Variables and the Dynamics of Supply and Demand; Mass & Senge on Alternative Tests for Selecting Model Variables; and Randers' very useful Guidelines for Model Conceptualization.)

Richardson GP (1986) Problems with causal-loop diagrams. Syst Dyn Rev 2(2):158–170 (Causal-loop diagrams cannot show stock-and-flow structure explicitly and can obscure important dynamics. Offers guidelines for proper use and interpretation of CLDs.)

Richardson GP, Pugh AL III (1981) Introduction to system dynamics modeling with DYNAMO. Productivity Press, Cambridge (Introductory text with excellent treatment of conceptualization, stocks and flows, formulation, and analysis. A good way to learn the DYNAMO simulation language as well.)

Roberts N, Andersen DF, Deal RM, Grant MS, Shaffer WA (1983) Introduction to computer simulation: A system dynamics modeling approach. Addison-Wesley, Reading (Easy-to-understand introductory text, complete with exercises.)

Sterman JD (1984) Appropriate Summary Statistics for Evaluating the Historical Fit of System Dynamics Models. Dynamica 10(2):51–66 (Describes the use of rigorous statistical tools for establishing model validity. Shows how Theil statistics can be used to assess goodness-of-fit in dynamic models.)

Wolstenholme EF (1990) System enquiry – A system dynamics approach. Wiley, Chichester (Describes a research methodology for building a system dynamics analysis. Emphasizes causal-loop diagramming, mapping of mental models, and other tools for qualitative system dynamics.)

Modeling for Learning: Systems Thinking and Organizational Learning

Kim D (1989) Learning laboratories: Designing a reflective learning environment. In: Milling PM, Zahn EOK (eds) Computer-based management of complex systems: International system dynamics conference. Springer, Berlin (A case-study of a process designed to convey dynamic insights to participants in a workshop setting designed around a management flight simulator game.)

Morecroft JDW (1988) System dynamics and microworlds for policymakers. Europ J Operat Res 35(3):301–320 (Describes the model-building tools available to managers and policymakers.)

Morecroft JDW, Sterman JD (eds) (1992) Modelling for Learning. Eur J Operat Res Special Issue 59(1) (17 papers describing models and methods to enhance learning, both for individuals and organizations. Covers elicitation and group process techniques, management flight simulators, and tools for capturing, representing, and simulating mental and formal models.)

Richmond B (1990) Systems thinking: A critical set of critical thinking skills for the 90's and beyond. In: Andersen DF, Richardson GP, Sterman JD (eds) International System Dynamics Conference, 1990 (Proposes a process and skill set to teach systems thinking. The process relies on learner-directed learning. The skill set includes general scientific reasoning and SD, supported by simulation.)

Senge PM (1990) Catalyzing systems thinking within organizations. In: Masarik F (ed) Advances in organization development. Ablex, Norwood (Presents a case study in which the use of system dynamics generated insights into a chronic business problem. Steps in generating, testing and disseminating a system dynamics model are described.)

Senge PM (1990) The fifth discipline: The art and practice of the learning organization. Doubleday Currency, New York (Introduces systems thinking as part of a wider approach to organizational learning. Conveys basic system structures to a non-technical business audience by means of anecdotes and archetypes.)

Decision Making

Morecroft JDW (1983) System dynamics: Portraying bounded rationality. Omega 11(2):131–142 (SD models represent decision making as boundedly rational. Reviews and contrasts the concept of bounded rationality as developed by Herbert Simon. Uses Forrester's Market Growth model to show how locally rational decision rules can interact to yield globally dysfunctional outcomes.)

Morecroft JDW (1985) Rationality in the Analysis of Behavioral Simulation Models. Manag Sci 31(7):900–916 (Shows how the intended rationality of decision rules in SD models can be assessed, and how one analyzes a simulation model and output to understand the assumed bounds on rationality in dynamic models. A model of salesforce effort allocation is used to illustrate.)

Sterman JD (1987) Expectation formation in behavioral simulation models. Behav Sci 32:190–211 (Proposes and tests a simple dynamic model of expectation formation in dynamic models (the TREND function). Shows how the TREND function explains a forty year history of inflation forecasts and several different types of long-term energy demand forecasts.)

Sterman JD (1989) Misperceptions of feedback in dynamic decision making. Organ Behav Hum Decis Process 43(3):301–335 (Describes an experiment with a simple economic system in which subjects systematically generate costly oscillations. Estimates decision rules to characterize subject behavior. Finds that people systematically ignore feedbacks, time delays, accumulations, and nonlinearities. These misperceptions of feedback lead to poor quality decisions when dynamic complexity is high.)

Sterman JD (1989) Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. Manag Sci 35(3):321–339 (Analyzes the results of the Beer Distribution Game. Misperceptions of feedback are found to cause poor performance in the beer game, as in other experiments. Estimates of the subjects' decision rules show they ignore time delays, accumulations, feedbacks, and nonlinearities.)

Selected Applications of SD

Abdel-Hamid TK, Madnick SE (1991) Software project dynamics: An integrated approach. Prentice Hall, Englewood Cliffs (Integrated SD model of the software development process. The model covers design, coding, reviewing, and quality assurance; these are integrated with resource planning, scheduling, and management of software projects. Includes full documentation, validation, and policy tests.)

Cooper KG (1980) Naval ship production: A claim settled and a framework built. Interfaces 10(6) (An SD model was used to quantify the causes of cost overruns in a large military shipbuilding project. One of the first and most successful applications of system dynamics to large-scale project management; initiated a long line of related project modeling work.)

Ford A, Bull M (1989) Using system dynamics for conservation policy analysis in the pacific northwest. Syst Dyn Rev 5(1):1–15 (Describes the use of an extensive SD model of electric power generation with endogenous demand. The model is used to evaluate strategies for conservation and new generation capacity. Includes discussion of implementation and integration of the SD model with other existing planning tools.)

Gardiner LK, Shreckengost RC (1987) A system dynamics model for estimating heroin imports into the United States. Syst Dyn Rev 3(1):8–27 (Describes how the CIA used SD to estimate the illegal importation of drugs to the US.)

Homer JB (1985) Worker burnout: A dynamic model with implications for prevention and control. Syst Dyn Rev 1(1):42–62 (Explains how knowledge workers can experience cycles of burnout through a simple system dynamics model. Avoiding burnout requires that one work at less than maximum capacity.)

Homer JB (1987) A diffusion model with application to evolving medical technologies. Technol Forecast Soc Chang 31(3):197–218 (Presents a generic model of the diffusion of new medical technologies. Case studies of the cardiac pacemaker and an antibiotic illustrate how the same model can explain the different diffusion dynamics of successful and unsuccessful technologies.)

Homer JB (1993) A system dynamics model of national cocaine prevalence. Syst Dyn Rev 9(1):49–78 (An excellent model of the interacting dynamics of addiction, policy-setting, and enforcement.)

Jensen KS, Mosekilde E, Holstein-Rathlou N (1985) Self-sustained oscillations and chaotic behaviour in kidney pressure regulation. In: Prigogine I, Sanglier M (eds) Laws of nature and human conduct. Taskforce of Research Information and Study on Science, Brussels (Presents a system dynamics model of the dynamics of rat kidneys. Experimental data show previously unexplained oscillations, sometimes chaotic. The model explains how these fluctuations arise. Excellent example of SD applied to physiology.)

Levin G, Hirsch GB, Roberts EB (1975) The persistent poppy: A computer-aided search for heroin policy. Ballinger, Cambridge (Examines the interactions within a community among drug users, the police and justice system, treatment agencies, and the citizens. Analyzes policies designed to restore the community's health.)

Levin G, Roberts EB, Hirsch GB, Kligler DS, Roberts N, Wilder JF (1976) The Dynamics of Human Service Delivery. Ballinger, Cambridge (Presents a generic theory of human service delivery, with case studies and examples drawn from mental health care, dental planning, elementary education, and outpatient care.)

Naill RF (1992) A system dynamics model for national energy policy planning. Syst Dyn Rev 8(1):1–19

Naill RF, Belanger S, Klinger A, Peterson E (1992) An analysis of the cost effectiveness of US energy policies to mitigate global warming. Syst Dyn Rev 8(2):111–128 (Reviews the 20 year history of the SD energy models used by the US Dept. of Energy to forecast and analyze policy options for national energy security, including the impact of US policies on global climate change.)

Sklar Reichelt K (1990) Halter marine: A case study of the dangers of litigation. (Working Paper No. D-4179). System Dynamics Group, Sloan School of Management, MIT, Cambridge (A case-study illustrating the use of system dynamics in litigation. Suitable for classroom teaching.)

Sturis J, Polonsky KS, Mosekilde E, Van Cauter E (1991) Computer model for mechanisms underlying ultradian oscillations of insulin and glucose. Am J Physiol 260(Endocrinol. Metab. 23):E801–E809 (New experimental data show that the human glucose/insulin system is inherently oscillatory. An SD model explains these dynamics. The model is validated against detailed physiological data.)

Cross-Fertilization and Comparative Methodology

Allen PM (1988) Dynamic models of evolving systems. Syst Dyn Rev 4(1–2):109–130 (Reviews approaches to nonlinear dynamics, self-organization, and evolution developed in the Brussels school by Prigogine, Allen, and others. Provides illustrations and examples.)

Kim DH (1990) Toward learning organizations: Integrating total quality control and systems thinking. (Working Paper No. D-4036). System Dynamics Group, Sloan School of Management, MIT, Cambridge (Argues that SD and Total Quality Management are complementary approaches to improvement and organizational learning. Systems thinking and modeling are needed to speed the improvement cycle for processes with long time delays.)

Meadows DH, Robinson JM (1985) The electronic oracle: Computer models and social decisions. Wiley (Comparative assessment of the underlying assumptions, boundary, limitations, and uses of different models, including optimization, simulation, and econometrics. Offers guidelines for assessing model assumptions, including ways to recognize the implicit biases of each modeling paradigm.)

Powers WT (1990) Control theory: A model of organisms. Syst Dyn

Rev 6(1):1–20 (An explicit feedback control perspective on perception and decision making in living organisms. Argues the behaviorist and cognitive paradigms have fundamentally misunderstood the concept of feedback. For Powers, feedback allows organisms to control perceptions by altering behavior.)

Radzicki MJ (1990) Methodologia oeconomiae et systematis dynamis. Syst Dyn Rev 6(2):123–147 (Surveys the institutionalist paradigm in economics and argues that system dynamics is compatible with the institutionalist perspective. The SD approach offers a means by which institutional theories can be formalized and tested.)

Sterman JD (1985) The growth of knowledge: Testing a theory of scientific revolutions with a formal model. Technol Forecast Soc Chang 28(2):93–122 (Presents a formal dynamic model of TS Kuhn's theory of scientific revolutions.)

Sterman JD (1988) A skeptic's guide to computer models. In: Grant L, Lanham MD (eds) Foresight and national decisions. University Press of America (Reviews different modeling methods and their underlying assumptions in nontechnical language. Provides a list of questions model users should ask to assess whether a model or method are appropriate to the problem.)

Other Themes: Pulling the Threads Together

Cooper K, Steinhurst W (eds) (1992) The system dynamics society bibliography. System Dynamics Society. Available from Julie Pugh, 49 Bedford Rd., Lincoln MA, USA 01773. (Lists over 3,000 system dynamics journal articles, books, conference proceedings and working papers. Available in computer-readable format and compatible with bibliographic software)

Meadows DH (1989) System dynamics meets the press. Syst Dyn Rev 5(1):68–80 (Reviews the history of encounters between SD and the media. Offers guidelines for effective communication to the public at large. Stresses the importance of communicating even the simplest system concepts.)

Meadows DH (1991) The global citizen. Island Press, Washington (A collection of Dana's syndicated newspaper columns applying system dynamics principles to problems of everyday life, from organic farming to the fall of the Soviet Union. Emphasizes environmental issues.)

Richardson GP (1991) Feedback thought in social science. University of Pennsylvania Press (Traces the history of the concept of feedback in the social sciences through two threads of thought – the cybernetic and feedback threads. System dynamics is placed in context in a readable and scholarly manner.)

Software

DYNAMO. Pugh-Roberts Associates, Cambridge MA. (The first widely-used computer language developed to simulate system dynamics models, DYNAMO is still in use, available for mainframes and PCs. Many of the models in the system dynamics literature were simulated in DYNAMO)

DYSMAP. University of Salford, UK (PC-based simulation language with syntax similar to DYNAMO. Includes optimization capability based on hill-climbing.)

Microworld Creator and S^4. Microworlds Inc., Cambridge MA (Easy to use environment for simulation and gaming. S^4, the 'industrial strength' version, supports arrays and includes diagnostics for analyzing behavior. Both Creator and S^4 support user-defined information displays and facilitate rapid development of management flight simulators.)

STELLA and ithink. High Performance Systems, Hanover NH. (User-friendly modeling software with full graphical interface. Models are entered graphically, at the level of the stock and flow diagram. Widely used in education from elementary school up; also used in research and practice.)

Vensim. Ventana Systems, Harvard MA. (Powerful simulation environment for SD models. Runs on workstations and PCs. Includes array capability and a wide range of features for analyzing model behavior.)

System Dynamics Philosophical Background and Underpinnings

Camilo Olaya
Universidad de Los Andes, Bogotá, Colombia

Article Outline

Glossary

Philosophy The reflection and study of our most basic assumptions – or the assumptions themselves.

Mental model A mental image of selected concepts and relationships of the world around us which we consider relevant for explaining the behavior of a particular system.

Presentationalism Synonymous of idealism. The view that material objects or external realities do not exist apart from our knowledge or consciousness of them.

Definition of the Subject

We all tend to take things for granted. Indeed it is a common place to judge formal models exclusively based on the technical grounds and on the logic with which those models were built without a proper reflection on the assumptions underlying those models. This omission is even more pressing in complexity and system science, since these areas represent a novel challenge for philosophers of science – e. g. see an overview in [34].

What is the idea of reality with which we work? What do we assume about human nature? What kind of knowledge do we pursue? What kind of knowledge do we obtain?

What is the scope of rational inquiry? What are the basis and the implications of our own reasoning methods? The identification of how philosophy has shaped the work of scientists – on a conscious or unconscious level – is essential for comprehending the implications, the limitations, and the scope of our very scientific practice. The lack of concern by scientists for these issues may explain many of their failures which has produced just a sort of inertial blindness that is easy to recognize in current scientific debates.

One of the strengths of system dynamics (abbreviated, SD) is that it leads us to make explicit our assumptions about the systems we deal with. This attitude, i. e. the importance of reflection upon our own assumptions, is also fundamental for the very development and practice of system dynamics. Many of the debates on different issues of every day scientific practice such as model conceptualization, formal model building, validation, policy design, etc. are informed and can be enlightened by the reflection on the philosophical background behind those processes. There are also various fundamental aspects of SD that are yet to be demarcated, e. g. the characterization of SD explanations. Furthermore, the ambiguity of the discussions found in large part of related literature, characterized by superficiality, confusion of terms, misdirected arguments, etc. only adds noise and it complicates the advance of a discipline. This article sketches and overview on some basic assumptions regarding the development and the practice of system dynamics. Various suggestions that help to integrate various debates are introduced and important clarifications are also indicated.

Introduction

The philosophical background and underpinnings of a discipline should have to do with its most basic universal assumptions. Such a discussion becomes difficult if we bear in mind that those assumptions are not necessarily shared by practitioners and researchers. Nevertheless, central premises can be identified which in turn can be related to important questions regarding philosophical concerns such as reality and knowledge.

The article is organized as follows. After this short introduction, the second sections develops an overview of the origins of system dynamics underlining fundamental aspects that formed what can be called the core of the discipline. This historical review highlights the initial interest, purposes and initial assumptions around the foundation of SD. With these elements the following sections introduce various philosophical issues that can be identified underlying system dynamics. Perhaps the central aspect is presentationalism, a stance associated with the notion of "mental models" which is central in SD; this is the topic of the fourth section. The following section makes a clarification on the controversial issue of positivism and relates presentationalism with knowledge. The sixth section summarizes the position of system dynamics regarding social theory. The seventh section presents the inquiry of explanation clarifying that in spite that SD involves causal models the nature of its kind of explanation can be found in the notion of mechanism. The eighth section introduces the implication of the use of computer simulation as a distinctive epistemology which is different from the traditional discourse in philosophy of science. The ninth section outlines future directions.

Before starting, a brief warning should be made: given the scope of this review and the limited space for covering very wide subjects then this article should be viewed as a broad introductory overview of the different topics. The cited literature has been selected as possible starting points for further inquiry.

System Dynamics

In order to address a "philosophical background" the first question naturally would be: What is system dynamics? Already this inquiry can be a matter of debate, e. g. [102]. Indeed SD has been labeled as a theory [23,49], a method [18,56,63,98,108], a methodology [81], a field of study [17,78], a tool [61], a paradigm, among other nouns. A natural starting point is the work of Jay Forrester, the founder of system dynamics. A brief historical review should help to grasp the very core we are looking for since it can show the initial motivations, assumptions, and purposes behind the development of SD.

Genesis of System Dynamics

Jay W. Forrester, member of the Sloan School of Management at MIT, was looking for linkages between engineering and management education given his background in feedback control systems and computers [28]. In 1956 he wrote a "note" to the Faculty Research Seminar, the first ever MIT "D-memo"; in this communication he sketched the worldview of what would be known as "system dynamics" [32].

He started with a strong criticism of economic models. The following were the central aspects: (i) their failure to reflect adequately the loop structures that make up economic systems; in particular this neglect leads to exclude inherent properties of closed loops such as resistance to change, accumulations and delays; (ii) The incapacity for including flows of goods, money, information, and labor, in one single interrelated model; (iii) The exclusion of changing mental attitudes that affect and explain economic processes; (iv) The use of linear equations for describing systems; (v) The restriction of building models constrained by the capacity for manipulating numerical data and solving the equations; (vi) The overconfidence in multiple regression analysis for obtaining coefficients for equations that define economic behavior; (vii) The lack of reflection and discussion on the very assumptions underlying every model preferring an emphasis on the logic with which the model is developed.

After delineating these points, Forrester then proceeded in the same note to highlight techniques that were largely underused at that time: servomechanisms, differential equations, and what he called "the art of simulation". Anchored on the mentioned assessment and on these developments Forrester conceived "a new avenue of attack for understanding the firm and the economy" ([28] p. 336) envisaging *a new kind of models* that would include aspects such as:

Dynamic structure: Detailed attention to the *sequences of actions* which occur in the system being studied and to the *forces* which trigger or temper such actions, with a particular concern on the controlling influences of lags and delays.

Information flows: Explicit recognition of information flow channels and information transformation with time and transmission.

Decision criteria: Re-examination of the proper decision criteria which must not be defined as depending only on current values of gross economic variables; instead, such criteria must be traced to the motivations, hopes, objectives and optimism of the people involved, including as well what he calls business man's intuition which "represents a disordered accumulation of basic insights into how people and social systems react. The hope for the future lies in generating an orderly arrangement of basic insights" ([28] p. 342).

Non-linear systems: Economic systems present most – if not all – of the time highly non-linear characteristics.

Differential equations: The behavior of economic systems should be better described by non-linear differential equations since they have been developed to describe delays, momentum, elasticity, reservoirs, and accelerations, which are better suited quantities for describing the economic world. In practice these equations would be handled as incremental difference equations in order to obtain numerical solutions.

Incremental changes in variables: To prefer the formulation of a model in terms of the motivations that cause incremental changes in a variable since the new value of a variable "can be found by solving the equations for its incremental change and then adding the change to the preceding full value of the variable" ([28] p. 344).

Model complexity: Much complex and complete models can be developed with these techniques.

Empirical solutions: It is useless to look for explicit unique or "correct" solutions; instead, these models provide diverse solutions according to the different assumptions about the model structure and the initial values of the variables.

Symbolism and correspondence with real counterparts: The possibility of having a pictorial representation – a flow diagram – whose processes of information, money, goods, and people, are moved, i. e. simulated, time-step-by-time-step from place to place.

Structure over coefficient accuracy: To prefer a structure in which we have confidence using intuitively estimated coefficients instead of using unlikely structures with accurately derived coefficients from statistical data.

A subsequent advance came in 1958 with an article entitled "Industrial Dynamics: A Major Breakthrough for Decision Makers" published in the *Harvard Business Review* [27]. In this article Forrester shaped the previous ideas with the concern that management should evolve from a highly fragmentized art to a profession capable of recognizing unified systems given that the task of management is to interrelate the flows of information, materials, labor, money, and capital equipment. He again emphasized features such as electronic data processing, decision making, simulation, feedback control, and information flows. These elements were presented as the cornerstones of the innovative industrial dynamics program at MIT.

Industrial Dynamics

The definitive breakthrough came in 1961 with his *magnus opus* "Industrial Dynamics" [24]. The main motivation behind was the development of a *science* for designing and controlling industrial systems, the quest for a management science. In particular he conceived it as "a method of system analysis for management." (p. 9). He stated:

> Industrial dynamics is the study of the information-feedback characteristics of industrial activity to show how organizational structure, amplification (in policies), and time delays (in decisions and actions) interact to influence the success of the enterprise (p. 13).

Forrester underlined four pillars for this new science: information-feedback control theory anchored on the concept of servomechanisms, the study of decision-making processes, an experimental approach to system analysis based on simulation, and the use of computers.

Coming from engineering, he had in mind models that deal with nonlinear dynamic systems whose purpose is to *design* new systems – as opposed to just *explain* systems; the models should show how changes in policies or structure will produce better or worse behavior. In order to accomplish this aim Forrester indicated that we should focus on understanding the characteristics of the system in hands (instead of looking for specific predictions) and on our assumptions about them. "We then have a means for tracing the implications of our assumptions" ([24] p. 55).

What should be included in a model? Forrester underlined that "there will be no such thing as *the* model of a social system, any more than there is *the* model of an aircraft ... the factors that must be included arise directly from the questions that are to be answered" ([24] p. 60). It is expected that these factors will include closed-loop information-feedback structures that give rise to so much of the interesting behavior. An important aspect of this new kind of models is symbolism and pictorial representations by means of flow diagrams with a special emphasis on correspondence: the model variables should correspond to those in the system being represented. In this book Forrester also demarcated the network structure of this new kind of models as made of four basic components: accumulations (levels), flows (rates), decision functions, and information channels; these networks trace cause-effect relationships which are described via mathematical formalization. He also discussed in detail how to represent delays and how to model decision processes which are particularly defined by general policies, i. e. rules that state how operation decisions are made converting information into action; this study of general policies explains the importance for SD of the examination of human decision-making processes. Another fundamental characteristics are continuous flows and aggregation: "grouping of individual events into classes ... Our interest in the model ... is from the viewpoint above the separate individual transactions" ([24] p. 65); it is assumed that different individual items are controlled by the same identical decision-function; this notion leads to aggregation which is a distinctive aspect of SD: "items controlled by sufficiently similar policies that depend on sufficiently equivalent information sources may be combined into a single channel" ([24] p. 109); the central criterion for such aggregation is the purpose of the model. Finally, model significance (or validity) rests on its suitability for a particular purpose which is motivated by the design of improved industrial and economic systems.

Principles of System Dynamics

The main concern in these initial steps were business operations. However, Forrester sketched a glance to a broader view in the final part of his book; there, he speaks of "system" dynamics since "the study of systems can provide a framework to unite subjects ... The dynamic model represents a system as broad as one chooses to describe" ([24] pp. 344, 346). Indeed, looking for a more general view he presents a section of "principles of systems structure":

> The principles to be discussed here all arise in the context of information-feedback systems. They are systems principles. They are not the principles of the management art such as have been taught in organization, production, and human relations courses. Because the principles apply to systems behavior, they do not fall into neat separate packages ... The concept of a system implies interaction and interdependence. In attempting to identify factors that are common to all systems, we must keep the essential indivisibility in mind ([24] pp. 347, 348).

Indeed, Forrester reaffirmed this general *systems* view in a follow-up book entitled *Principles of Systems* [25] published in 1968. He gives a basic definition of system as "a grouping of parts that operate together for a common purpose" ([25] p. 1-1). Moreover, he suggests that the way to organize knowledge is with this idea of system which are represented by the models we develop:

> A structure (theory) is essential if we are to effectively interrelate and interpret our observations in any field of knowledge ... Without an organizing structure, knowledge is a mere collection of observations, practices and conflicting incidents ... A model is s substitute for an object or system ... Any set of rules and relationships that describe something is a model of that thing. In this sense, all of our thinking depends on models. Our mental processes use concepts which we manipulate into new arrangements. These concepts are not, in fact, the real system that they represent. The mental concepts are abstractions based on our experience. This experience has been filtered and modified by our individual perception and organization processes to produce our mental models that represent the world around us ([25] pp. 1-2, 3-1).

The previous statement summarizes core assumptions for SD. It points at the central notion of *mental model* and it emphasizes that these models are models of *systems*. In fact, in an article published in 1968 in *Management*

Science, Forrester [28] underlined that this application of feedback concepts to social systems was evolving toward a theory of structure in systems with a particular goal of policy design. Specifically, "industrial dynamics is a philosophy of structure in systems. It is also gradually becoming a body of principles that relate structure to behavior" ([28] p. 141). Two fundamental variables are identified: levels and rates, and the basic structural element is the feedback loop: "every decision is responsive to the existing condition of the system and influences that condition" ([28] p. 143). And, as stated above, the structure is an aid to organize knowledge in a particular situation given a particular purpose which is motivated by the pursue of explanation and policy design.

This historical review has accentuated central aspects of the foundation of SD. As a summary, this science initially envisaged by Forrester for designing systems is known nowadays as system dynamics. The models developed in SD have distinctive characteristics: dynamic structures, information flows, study of decision criteria, non-linearity, difference equations, symbolism and correspondence, and emphasis on confidence based on the structure of the model. The practice of this science is anchored on: the concept of servomechanism, the study of decision-making processes, the embrace of an experimental approach, and the use of computers for building formal models. These models are models of systems and the main goal is to help to organize our knowledge – which can be seen as arranged in mental models – so as to enhance learning processes and systems design on concrete settings and under specific purposes.

With these elements in mind, it is now possible to proceed with a brief discussion on important aspects regarding assumptions on reality and knowledge that can be recognized behind these premises and the practice of SD.

"Real" World and Presentationalism

The traditional distinctions of *ontology*, *epistemology*, and *methodology*, form the habitual framework to drive a philosophical discussion. But the isolation of these issues may not be the most adequate or clear strategy. Given the interrelated nature of such categories and the misleading discussion on those terms which is present in a large part of organization and management science literature – part of this confusion will be exposed and clarified below – then a different plan will be used to develop the rest of this article. An instinctive option is to pick up significant issues and to relate them with what we can identify as part of the core of SD.

Where to start? Assumptions concerning a "real" world can be a first step to take since the traditional debate developed through the years around the claim of building models of "social systems" has fueled part of the discussions in systems science. As most of examinations on philosophical matters the debate has been permeated by a confusion originated in terms and words without the proper examination of the topic. However, this short revision is useful for opening the assessment of the premises of SD regarding a real world.

A prominent example is the criticism made by influential commentators who state that SD models represent an assumed "objective" real world [49]. This kind of critique usually labels SD as a "hard" approach, e. g. [48], meaning with such a term models of an assumed objective machine-like world – and habitually including and inverting the meaning of the term "positivism" – a mistaken assessment that still can be seen nowadays, e. g. [16]. This type of comments can be illustrated with the following quote from the work of Flood and Jackson [23]:

> System dynamics models still center on capturing the structure of the "real world" … the underlying assumption of SD that there is an external world made up of systems the structure of which can be grasped using models built upon feedback processes … Because intentions derive from inside social systems, from the conscious human actors which constitute them, many possible appreciations of the nature and purpose of particular social systems are possible. SD simply does not deal with the innate subjectivity of human beings … In essence the argument is that social systems cannot be studied, in the way of system dynamics, objectively from the outside (pp. 78, 79–80).

However, that is not what SD looks for. The mentioned emphasis on the examination of human decision-making processes and on the assumptions behind, the notion of mental model, the fact that model significance rests on its suitability for a particular purpose, among other aspects, should suffice to illustrate the point. Already Forrester in 1961 [24] emphasized: "a model can be useful if it represents only what we *believe* to be the nature of the system under study … we are forced to commit ourselves on what we believe is the relative importance of various factors. We shall discover inconsistencies in our basic assumptions … Thorough any of these we learn" (pp. 57–58, emphasis original). This should be enough to discard the assessment made by Flood and Jackson, and similar critics, who mistakenly placed SD in the terrain of a sort of naive realism as depicted in the quote above; this point is extensively clarified by Lane [55,56]. More importantly, this argument is helpful to introduce the discussion about the nature of SD models, the assumptions behind these models, and their relation with a "reality". The notion that drives these matters has been labeled in SD literature as "mental model" and even though this concept is not free of debate [21] it is placed at the core of the discipline.

The idea of mental model was already addressed by Forrester, as it has been indicated; he demarcated it in 1970 as the mental image of selected concepts and relationships of the world around us that we carry in our heads [26]; furthermore, "the mental model is fuzzy. It is incomplete. It is imprecisely stated … [It] changes with time" (p. 213). Doyle and Ford [22], looking for a consensus on this subject, propose as a definition: "a relatively enduring and accessible [conscious], but limited, internal conceptual representation of an external system (historical, existing or projected) whose structure is analogous to the perceived structure of that system" (p. 411). Sterman [98] summarizes what the expression "mental model" refers to: "our beliefs about the networks of causes and effects that describe how a system operates, along with the boundary of the model (which variables are included and which are excluded) and the time horizon we consider relevant … Most of us do not appreciate the ubiquity and invisibility of mental models, instead believing naively that our senses reveal the world as it is. On the contrary, our world is actively constructed (modeled) by our senses and brain" (p. 16–17). It should be noticed that these mental models may refer as well to planned or desired systems existing in the mind of the modeler [54]. The ultimate goal of system dynamics is to enhance our learning processes by testing and improving our mental models in a way that becomes consistent with the complexity of the systems that we face and design everyday [98].

Then, what is a SD model as related to some "real" world? Or how is it related to the notion of mental models? The usual option to answer these questions is to frame the discussion in the debate realism/anti-realism. This examination is even more relevant considering that some criticism labels SD as anchored on "realism". A brief clarification follows.

Realism? Anti-Realism?

Typical of latest debates in history of science concerns the dispute between the so-called *realism*, i. e. theories are true *or* false as descriptions of the world, and *instrumentalism*, i. e. theories are more or less adequate. This latter position is closer to pragmatism: theories are just instruments to systematize – and for many philosophers predict – observations, but theories are not claims about the world; or they can be also subjected to linguistic frameworks but still always truth-value-less. An overview of this debate is made by Leplin [60]. Yet, here it should be clarified that both positions and those definitions – habitually taken by historians of science – are two sides of the same coin. In short, both are forms of idealism. This assessment will be commented next.

On the one hand, scientific "realism" is usually defended using the method of abduction as source of knowledge, i. e. inference to the best explanation, which is just a form of induction and hence it is confirmationist, ultimately relativistic. It is not realism at all; it is exactly the opposite, i. e. idealism. For instance, take the influential ideas of Sellars [91] who defends an illustrative traditional position of scientific realism holding the source of knowledge on observation and relying on justification by induction for building theoretical frameworks; he emphasizes: "Laws in question are stipulated to be inductively established in the observation framework" (p. 313), the theories are then refined by empirical generalizations and what he calls further "injections" of images of the theory into the observational framework, in a typical process of instructional correction. A further refined and formal model of such a framework (proposed by Friedmann, holding a model–submodel relation) is commented by Morrison [66] who still holds, nevertheless, the search for truth and justification as support for his criticism and the necessity of confirmation [67]; see also for example the paper of Kukla [52] with a criticism to the ideas of Friedmann, yet also supported also on confirmation. In short, this "realism" is just idealism as we know it. Similar "realist" positions abound on the literature of the history of science. E.g. Smith [95] postulates a realist stance based on "common sense" but, since for him "science begins with observations" (p. 53), his realism ends up, indeed, trapped in sensations, i. e. idealism. Quantum mechanics does not escape the debate, e. g. arguing in favor of realism within the subjective theory of probability (the Copenhagen interpretation, Bohr, Heisenberg, etc.), Dickson [20] stands on the verification criterion for discussing on what he calls "quantum realism"; naturally such base can not succeed, which is anyway best self-explanatorily reflected in the first part of the title of his paper "An Empirical Reply to Empiricism"; simply there is no such reply. Another case is the criticism of Brown [13] to the deterministic notion of "realism" that Cherniak discards (who supports it on computer simulation of finite agents); the criticism of Brown shows a Galilean notion of realism, i. e. achievable true descriptions of the world via laws, but in a very "complex world" and thus, for him, inaccessible to agents with limited cognitive capacity; this Galilean view is found in large part of complexity science. In other attempts, for instance Schlagel [90] defends "contextualistic realism" which ends up in relativism: elements of the world have a conditional status relative to contexts and conditions, the existents are real relative to the particular structures and contexts on which they depend; his proposal is just idealism and can be better described as a sort of multi-phenomenalism, indeed relativism. The diverse "realist" positions *actually rooted on non-negotiable empiricism* seem endless; e. g. further examples are found in [1,59,75], and in most papers with the expression "scientific realism" in the abstract.

On the other hand, let us consider a supposedly opposite position, for instance instrumentalism. Instrumentalism actually ends up in relativism as well, e. g. with respect to the points of view where instruments can be applied; at the end the notions of "applicability" or "adequacy" become reference frameworks for establishing partial truths. Newton-Smith [68] proposes a conciliatory position, he calls it "modest realism" which ends up indeed in a sort of "moderate" relativism; a catalog of various scientific realisms can be found in that paper too, all of them addressing still the question of how such truths about a real world can be sustained (justified). In short, instrumentalists declare the dependence on observations, i. e. idealism, and thus this position is in fact sharing assumptions with the alleged "realists". Leplin [60] summarizes: "some theory can be reduced to observation by defining or translating theoretical terms into terms that describe observable conditions. The remainder must be construed instrumentally" (p. 394).

The "realism" vs. instrumentalism debate is futile and yet it is perhaps one of the core discussions in philosophy of science. But unlike these influential historians of science, the presented dispute tends to be dismissed by several professional philosophers who see it as self-serving and unsophisticated; Fuller [33] underlines this assessment though he also illustrates the consequences of leaving such pointlessness discussions to endure. In this case, two seemingly unaware idealist factions argue about the best way to establish positive knowledge, e. g. either with a supposedly "true" description of the world (more precisely: *phenomena*, subjectivism) or by pragmatism (and again: *phenomena*, relativism). This shared *idealism* is the matter of the next section.

Presentationalism

A broader debate can be assessed framed in the opposition between *realism* and *idealism*, see e. g. [72] and [77]. It will be shown that SD rests on the broad stance known as idealism, i. e. presentationalism – this latter term is preferred here just for clarity [10].

Firstly, it should be commented the widely inverted and misleading use of both terms. And there is good company. Blackmore [10] shows various celebrities that became misusers of the expressions in question such as the former president of the American History of Science Society and prominent Harvard University professor, Erwin Hiebert, and Sir Russell Brain, outstanding neurologist and former president of the British Association for the Advancement of Science. Add the seemingly customary tendency to quote references in second hand without inspecting direct sources and a few decades later we have the terms used in exactly their opposite original sense in journals and books. Blackmore pictures the situation:

> Like-minded 'empiricists' have restricted what they understand by the term to what idealistic philosophers of science *have wanted them to understand by it*. And since the term 'realism' sounds good to 'tough-minded empiricists' and since idealistic philosophers such as Hume, Comte, Schuppe, and Mach and their recent successors *scarcely if ever have admitted their idealism*, many scientists and historians of science have let themselves be seduced into reversing the normal epistemological definitions of 'realism' and 'idealism'. Even worse, respected scholars such as Stillman Drake, perhaps our most outstanding authority on the manuscripts of Galileo, and Larry Laudan, a young, energetic, and much published commentator on Mach and 'empiricism', have allowed themselves to become advocates of hopelessly naive 'non-philosophical' positions. Drake is sure that Galileo held no philosophical position, or that if he did, it had no effect on his scientific work. Laudan is equally positive on the basis of Mach's written comments that his phenomenalistic epistemology had no influence on Mach's 'empirical' methodology of science. The simplicity of their views can party be explained by their tendency to understand by the term 'philosophy', not one's most basic universal assumptions, *but expressed talk about speculative matters*. Similarly, many 'materialists' who feel sure that the physical world is directly given in experience, and who accept the idealist Kant's distinction between 'science' and 'metaphysics', are convinced that anyone who identifies the physical world with what is *beyond* immediate experience is an 'idealist' and 'metaphysician' and (following Wittgenstein) 'is merely uttering nonsense' (p. 131 in [10]).

In order to clarify, it is appropriate to underscore the attitude behind an empiricist epistemology. A major defining posture is what has been labeled as *idealism*, given the natural disbelief of a world beyond senses, which is the pillar of an empiricist epistemology. The term "idealism" comes from the "idea" of Bishop Berkeley, who took physical objects as "ideas" which included sensations and thoughts: "It is evident to any one who takes a survey of the objects of human knowledge, that they are either ideas actually imprinted on the senses, or else such as are perceived by attending to the passions and operations of the mind, or lastly ideas formed by help of memory and imagination, either compounding, dividing, or barely representing those originally perceived in the aforesaid ways." [8]. This idealism relies entirely on senses and mind-dependent worlds since consequently sense-data were the only things of whose existence our perceptions could assure us, and that to be known is to be 'in' a mind, and therefore to be mental. Berkeley, therefore, concluded that nothing can ever be known except what is in some mind, and that whatever is known without being in my mind must be in some other mind [85]. Hunter [47] summarizes:

> As a result of their constraints on knowledge and meaning, empiricists tend to be skeptical of necessary truths that are independent of mind and language, and of putative eternal abstract entities (p. 110).

This idealism can be equally identified with terms such as 'phenomenalism', 'neutral monism', or 'subjective idealism' (e. g. [10]), or presentationalism. In other words, "anything in time or space, anything than can be known by the human mind, is phenomenal" (p. 146 in [15]).

Taking this posture to the context of science, Bartley [4] provides the implications: "Presentationalists see the subject matter of science not as an external reality independent of sensation. The subject matter of science is our sensory perceptions. The collectivity of these sensations is renamed 'nature' … The aim of science is seen not as the description and explanation of that independent external reality but as the efficient computation of perceptions … [It] became the dominant twentieth-century *philosophy* of physics" (p. 11, 16, emphasis original). In general this position is the pillar of the prevalent conception of science which has been fueled by physics (for instance in the interpretation of quantum mechanics of Bohr and Heisenberg) and backed by influential names like Mach, Russell, Wittgenstein, Ayer, Lewis, Carnap, etc.

To appreciate the contrast, a standard definition of *realism* can be:

> 'Realism' … is used for the view that material objects exist externally to us and independently of our sense experience. Realism is thus opposed to idealism, which holds that no such material objects or external realities exist apart from our knowledge or consciousness of them, the whole universe thus being dependent on the mind or in some sense mental. It also clashes with phenomenalism, which, while avoiding much idealist metaphysics, would deny that material objects exist except as groups or sequences of sensa, actual and possible (p. 126 in [10]).

Apart from the particular emphasis on material objects (as opposed to Berkeley's *ideas*), another point is that realism defends a cosmocentric thesis opposed to the anthropocentric view in the discussions of the alleged "realists" of science presented earlier; in the latter case the observer-centered learning process is fundamental, and it is carried on via induction looking for acquiring positive, verifiable, and true knowledge – or justified true belief. Moreover, let us recall that historians of science denote with the term "realism" just the concern with supposed true descriptions of the world. But in fact *realism* does not imply that knowledge is achievable, it does not imply that the world is a perfect clock, it does not imply determinism, it does not imply that there can be correspondence between theories and such real world; these are different affirmations that unfortunately seem to be muddled inside the same bag. One thing is to assume a real world beyond senses. But a different inquiry is the character we ascribe to it. Another different issue is the role we assume for our senses. Another very different concern is the question of knowability, etc. Rescher [77] illustrates typical examples of the confusing use of terms in literature: "The three positions to the effect that real things just exactly are things as *philosophy* or as *science* or as '*commonsense*' takes them to be – positions generally designated as *scholastic*, *scientific* and *naive* realism, respectively – are in fact versions of epistemic idealism exactly because they see reals as inherently knowable and do not contemplate mind-transcendence for the real" (p. 187).

Coming back to presentationalism, this position then assumes that we are imprinted by the environment, and we call to this *impressions* "knowledge". Given the limitation of our senses then presentationalism postulates that nothing more can be known; and thus, such assumption is used to construct the world in our minds: for a presentationalist, strictly speaking, the world is not re-presented since we do not have access to it, the world is just what is *presented* to our senses: the world as we experience it happens to be the world itself; and, since anything that can be known by the human mind is, then, phenomenal (sensations, etc.), therefore knowledge strictly depends on – and is source in – what is sensed, e. g. observed. Hence knowledge needs to be justified in order to avoid error; and yet, the only existent knowledge is the imperfect evidence sourced in sensation and, nevertheless, a foundation that can be justified is pursued. This is the popular plan we have come up with, so far, to try to avoid the destruction of empiricism made by Hume. The picture can be summarized:

> Almost all traditional epistemologies are Lamarckian in their accounts of the growth of knowledge. This is conspicuously true of presentationalism, almost all adherents to which maintain an inductivist, justificationalist account of knowledge growth, according to which knowledge is constructed out of sensations (as building blocks or elements) by a relative passive process of combination, accumulation, repetition, and induction (p. 25 in [4]).

This position is identified with *idealism* and most popularly associated with *epistemological empiricism*, with all its assumptions and its consequent scientific method.

The last point to underline is that this epistemology has subordinated ontology. A remarkable illustration of this type of problems was already made by Bowman [11]: "The result … is the more or less deliberate abnegation of a genuine epistemology and the substitution for it of a highly formal logic. Hence the paradox illustrated equally in the case of Plato and, recently, of Mr Russell, of a radical empiricism (expressed in Plato's Protagorean theory of sensation and in Russell's subjectivism) subsisting side by side with the extremist rationalism. Such a dualistic position is the despair of philosophy, and indicates a failure in the synthetic work of thought" (p. 485). This despair is easy to recognize in current science which presents a contradictory ontological position whose difficulty is found in its subsumption under a radical epistemology. Indeed extreme empiricism, i. e. presentationalism, has become the metaphysics, i. e. the theory of reality, of our science. The debates on "scientific realism" presented earlier picture this failure. On this particular subject the different use of terms in literature is a source of confusion; but here the terms have been inverted by historians of science and scientists; and beyond a semantic confusion this has brought a narrow conception and a very restricted examination of epistemological assumptions. In short, the subjectivism of Descartes and Kant – or more precisely, Kantian idealism – is what now is labeled as "realism", e. g. *everything* has become "phenomena". As a matter of fact common expressions like "objective phenomena" or "real phenomena" uncover an idealistic position where the "objective" or "real" are just *phenomena*. Indeed, the so-called scientific "realism" commented above is nothing more than *a sort of empiricism* driven by Hume and Kant. This "realism" is just ontology overlapped by epistemology (*idealism*). More precisely, regarding Kant, let us recall his "Transcendental Idealism" which was the common answer of Kant when he was accused of idealist, e. g. see [93], denying to be a "dogmatic idealist" (in the Berkeley sense; see e. g. [101]); a full discussion of this failed defence of Kant is made by Guyer [37]. In particular Turbayne [101] defends Kant when he was accused of having misinterpreted (or even completely having misunderstood) Berkeley's idealism; yet, Turbayne's conclusion summarizes the Kantian idealism (and its ambiguity) unmistakably: "The Kantian antidote to this is not the a priori nature of space, but *its reality or subjectivity*, which assimilates space and its contents into the realm of ideas, and thus *prevents* illusion" (p. 243, emphases added). Regarding consequences, perhaps the best summary is the radical idealist position of Mach – who denied even the existence of atoms since they cannot be observed. Kant seems to be taken for granted without a proper reflection on his position.

These few points were commented since this debate is the major informer of the method and the assumptions of management science, organization science, and social science in general, places where SD has its roots. More important, by making this clarifications then a clear ground for SD can be envisaged. It can be affirmed that SD has been mistakenly labeled as "realist" by many commentators alluding the alleged aim of building "true" descriptive theories of the world, and using the misleading definitions of historians of science. But also from the discussion above it should be clear that SD safely rests on presentationalism. Moreover, the identified presentationalist stance known as "instrumentalism", and in general the so-called anti-realists positions (independent of the inverted use of the term), fit to the SD worldview: the abstractions from our experience are arranged in mental models which form knowledge that we want to improve in order to make better decisions. The SD models built for achieving this goal are judged against their adequacy and suitability for a particular purpose; these models are not claims about the world but instruments for systematizing observations and for boosting learning processes using experimentation via simulation.

The Discussions on Positivism: Presentationalism and Knowledge

An issue previously mentioned which is closely connected with presentationalism is relativism and positivism. Since positivism usually – and mistakenly – is pejoratively associated with a supposed objective representation of reality, then an important clarification is needed. In short: positivism is consistent with presentationalism and with relativism as well.

Presentationalism Brings Positivism

Blackmore [10] reminds that strictly speaking neither "rationalism" nor "empiricism" are properly epistemological terms at all; the entrenched idealism has led just to this narrowed identification. For instance, usually the term "empiricism" is synonymous of knowledge sourced in observation, i. e. in a restricted epistemological context, but such popular narrowness is inaccurate. Blackmore remarks: "Granted, that if one *means* by 'empiricism' not just an extensive and careful concern with empirical evidence but *restricting* reference or knowledge or both to sensory appearances, then there are indeed epistemological implications. One has become an epistemological phenomenalist or subjective idealist, or if you will, a positivist" (p. 130, emphases original). This clarification is needed for two reasons; on the one hand, as it was stated, SD has been labeled as "positivist" but the critics take this term as a sort of naive realism; the confusion is patent once we realize that positivism actually is a consequence of idealism, the opposite doctrine of realism. On the other hand, since SD is better identified with idealism then a sort of positivism can be also associated with it, but not the sort of "positivism" that the critics have in mind but the authentic positivism; and yet we will see that with the use of simulation positivism does not necessarily fit either.

The case of management science is a good example regarding the discussion on relativism and positivism. Let us consider the traditional and unfortunate sharp division between "hard vs. soft" which also takes the form "quantitative vs. qualitative". However, this discussion is misleading as well. It is not difficult to find researchers that claim to be anti-positivists but being themselves grounded on positivism (e. g. empirical observation, verification, induction, etc.) without noticing the contradiction. A good example is the claimed opposition between positivism and phenomenology. Yet, phenomenology is authentic positivism when is committed to evidence – Husserl himself underlines this aspect – see e. g. [94]. Indeed it is easy to appreciate an inverted use of the term "positivism" in literature; in management science this is a favored and widespread practice where so-called anti-positivists do not notice their positivism. The fact is that anthropocentrism is the ground in our most influential epistemologies that recognize the obvious imperfections of our sensorial apparatus but, nevertheless, rely knowledge on sensation (observation, etc.), that is, positivism, which is nothing more that our anxiety to confirm and validate, i. e. justify, our "imperfect" knowledge, e. g. empirically. This is a simplification of highly loaded terms; clarifications and further discussions can be found elsewhere, e. g. regarding positivism see [9,97,100].

Positivism is Anchored on Justification

Within a presentationalist worldview the search for confirmation and verification is nothing less than the search for justification of knowledge where the intellectual authority lies in sense experience. From a presentationalist account it is straightforward to have a justificationist approach for confirming and verifying theories. Following Bartley [4]:

> Preoccupied with the avoidance of error, they suppose that, in order to avoid error, they must make no utterances that cannot be justified by – i. e., derived from – the evidence available. Yet sense perception seems to be the only available evidence … The claim that there is an external world *in addition* to the evidence is a claim going *beyond* the evidence. Hence, claims about such realms are unjustifiable. Crucial to the presentationalist argument are, then, two things: the desire to give a firm foundation or justification to the tenets of science, and the construal of sense experience as the incorrigible source of all knowledge (pp. 12–13, emphases original).

In fact justification philosophy taken as the search for epistemic 'authorities' has been the dominant style of western philosophy looking for "well-grounded" knowledge. For instance in the customary view of knowledge as *justified true belief*, e. g. in the sense of Russell [86]) – as the result of systematic analysis "of our sensory experience of a knowable external reality" (p. 47 in [96]). Within this popular position the central problem of epistemology – as succinctly formulated by Radnitzky [76] – becomes:

> "When is it rational or, so to speak consistent with one's intellectual integrity, to *accept* a particular position?" The formulation suggests the direction in which the answer is to be sought: "When concerned with a statement, a theory, etc., accept those and only those statements, theories, etc., which not only are true but whose truth has been established" (p. 282, emphasis original).

The goal of justification is usually entrenched within the method of induction where every new repeated observation is a confirmation that validates – justifies – the theoretical statement. Even with weaker conditions the way of reasoning is the same, for instance within the ideas of Ayer where strict verifiability is seen as a too rigid criterion – he introduces *confirmability to some degree*, instead of complete and conclusive verifiability (see [88]); justification is still pursued. The appeal of justification can be explained because it looks for avoiding relativism (inherently attached to presentationalism) since not all positions are equally good or bad and it suggests to look for something beyond blind belief [76]. Though it is not the only option, nevertheless, it is the most common view of science; the concerns on validation, justification, verification, confirmation, and generalization, are part of this popular and influential view. Here the observer is the fixed point of reference. In short, within a justificationist logic, it is rational to accept only those positions that have been justified according to the rational authority which in the case of presentationalism is sense experience, consistent then with the highly influential ideas of Locke, Berkeley, Hume, Mach, Carnap that have shaped our prevalent view of science [4].

Justification in System Dynamics

Turning back to SD, it must be recalled the role of mental models whose characteristic nature of "abstractions based on experience" can be better assessed within a presentationalist stance. Here justification has also a place. How is this knowledge justified? Already Forrester [24] emphasized, within the debate of model validation, that, "knowledge of all forms can be brought to bear on forming an opinion of whether or not a model is suitable to its particular purpose" (p. 129). Therefore, Forrester [25] also emphasized that "we can never prove that any model is an exact representation of 'reality' … Models are then to be judged, not on an absolute scale that condemns them for failure to be perfect, but on a relative scale that approves them if they succeed in clarifying our knowledge" (p. 3–4). This sort of relativism will be addressed next.

With the aim of placing this stance within the discussion of history of science, Barlas and Carpenter [3] addressed the "philosophical roots of model validation" associating SD with what they called a "relativist philosophy of science". In this view justification is pursued: such a knowledge is seen as socially, culturally and historically dependent and it becomes socially justified belief. Here "a valid model is assumed to be only one of many possible ways of describing a real situation … for every model carries in it the modeler's world view … validation is a matter of social conversation" (p. 157). Hence, confirmation and verification are pursued through a social process relative to a frame of reference. This sort of moderate relativism was later criticized by Vásquez, Liz and Aracil [103] for whom such relativism is unacceptable in spite of their recognition that there is no privileged single model (or set of models); since they are also concerned with epistemological justification these authors present Putnam's internal "realism" as a more adequate way to conceptualize the type of knowledge consistent with the assumptions about reality held by SD practitioners; in short, these authors underline that mental models are the source of knowledge – and its justification – helping to select the structures that must be assumed as working in real systems; here knowledge is taken as internal to the conceptual scheme of SD. Since there can be many models for a given situation, the authors argue that this framework gives the possibility of convergence as a result of "the strong interactive character of mental models" (p. 34) recognizing that in any case SD modeling is a process of revision and adjustment. With this proposal these authors seek to achieve justification and some realistic representational content in spite of the plurality of alternative SD models available for a specific situation. It is easy to see that this proposal is still relativistic, in this case knowledge is relative to the mental models and the conceptual scheme – though the mentioned authors would not agree since for them there is a "reality" given by the internal representational schema, in this case the mental models of the modelers.

The fail to recognize presentationalism as the epistemology which is driving ontology is at the root of the presented discussions. This is a distinctive trait present in large part of the philosophy of science literature. However, the characteristic problem of mistaking positivism with a sort of supposed "objectivism" is also present in this discussion – in fact, Barlas and Carpenter, following the misdirecting literature on the subject, argue that the relativist philosophy that they defend rejects positivism; Vásquez, Liz and Aracil also follow the same inertia. However, these theses, which seek for confirmation, verification, justification, reflect the search for *positive* knowledge within an idealist epistemology. In the first case, knowledge is confirmed and accepted through social interaction and it is relative to a context. In the second case, knowledge is justified on mental models. To appreciate a genuine contrasting position see an anti-justificationist approach to validation in computer simulation in [51]; the core of anti-justificationism can be found in the work of Popper, e. g. [73,74].

System Dynamics as Presentationalist

It should be clear that the assumptions on SD reject naive realism; the models are not supposed to be accurate and corresponding descriptions of an external true reality – and furthermore, this is not what the term "positivism" refers to. In any case validation in SD does not mean a supposed "positive proof" or to assess the development of "true models" of the world. On the contrary, SD aims at enhancing our ways of reasoning, it emphasizes a process of learning so as to consistently improve our mental models which are product of our experience and the operations of our mind. Hence SD is closer to presentationalism. Knowledge can be socially justified and our mental models can be enhanced. Moreover, a particular emphasis on the modeling *process* is also underlined – see e. g. [29,44,89] and it will be commented in the eight section devoted to simulation.

A Brief Note Regarding Social Theory

Another important discussion is related to the theories about the "social world" that are supposedly held in SD, i. e. the social theory behind SD if any. Part of the misguided debate is explained by the widespread use of the traditional framework of Burrell and Morgan [14] whose oversimplification of social science in four paradigms has deviated major issues covering important topics under a too practical and inadequate schema – e. g. see a criticism in Deetz [19]. Jackson [49] provides a summary of such usual misconstruction:

> System dynamics … is essentially functionalist in nature. It sees system structure as the determining force behind system behavior … If humans are free to construct social systems as they wish, what determining influence does system structure have? … This tension between determinism and free will is unresolved (p. 39).

It should suffice to recall from the discussion above that the *aggregate* approach of SD is not a theory of human behavior; SD is not concerned with *individual* action. Furthermore, it does not assume that a structure, of any kind, determines human behavior either, i. e. the sort of determinism that Burrell and Morgan [14] oppose to "free will" and in the lines of the already vague term known as "structuralism" – see an early clarification of the problems of such type of oversimplification in [84]. In any case, this sort of criticism has been answered and clarified by Lane [56] who has underlined the main point: SD is concerned with aggregate social phenomena and not with individual meaningful actions [55]. Moreover, system dynamics does not propose invariant causal laws, as Lane [56] also concludes: "The only universal law/theory on offer is a grand methodological, or structural theory, associated with a representation scheme … it does not attract the determinism-related criticisms attached to grand theory in the sense of Parsons and Mills" (p. 111). In [57] Lane proposes to link system dynamics with a different framework: agency/structure theories.

Servomechanism

Part of the confusion is because of the misunderstanding by various commentators of the notion of feedback that underlies SD. This point has been also a source of misconception given the use of feedback in theories of control applied to social systems, e. g. cybernetics. Consider for example the following comment by Flood and Jackson [23]:

> The attempt of SD thinkers to model external reality is misguided … The emphasis placed on "structure" as the means of revealing knowledge about the optimal behavior of systems cannot be accepted … SD modelers using feedforward control appear to believe that there are optimal future states that we should steer systems towards" (pp. 80, 81).

Such criticism apparently is associated with the notion of feedback used in cybernetics. However SD does not pursue optimization, let alone by studying "knowledge revealing" structures in order to achieve supposed optimal behavior patterns. Instead, the goal is to have a better understanding of feedback structures in order to enhance decision-making and policy design. This point has been also addressed by Lane [53,55]. The central clarification of this issue has been made by Richardson [78] who distinguishes two different threads in the development of the concept of feedback in the social sciences: the cybernetics thread and the servomechanisms thread. The failure in noticing these two different lines of thought has produced various misconceptions regarding the notion of feedback in SD, a concept that has been shown as one of its building blocks. On the one hand, the cybernetic conception of feedback is defined in terms of input and output, it is limited only to loops of negative polarity which in turn are conceived as the mechanisms of control – and hence there is a particular interest in goal seeking and goal formulation given the concern in cybernetics for achieving adaptive behavior via directed processes and homeostatic mechanisms; feedback mechanisms guide this pursue of viable behavior which is carried by goal-seeking processes. On the other hand, in SD, coming from the servomechanisms thread, feedback loops are taken as intrinsic parts of the system (and not just as mechanisms of control), it includes also loops of positive polarity, and such feedback structures are seen as responsible for counterintuitive behaviors and policy resistance in social systems; here the analysis is directed toward policy design.

Explanation and Mechanism

The next interesting question would be how we can achieve better understanding and better policy design by enhancing our mental models. How can we characterize this type of knowledge?

A solid account of *explanation* should be placed at the heart of any scientific activity. The general inquiry about [scientific] explanation has to do with "learning how the process of doing science facilitates understanding, and what type(s) of understanding science provides" ([7] p. 307). In a very intuitive way, a first approach to explanation might be associated plainly with removing puzzlement [6]. It is also common to affirm that an explanation aims to answer queries of *why* in order to provide understanding [87]. Yet, to characterize such idea is a major and open unresolved question in philosophy of science [69]. The notion of causality has traditionally played a central role; this view has pervaded most of scientific research where theory development and explanations are essentially conceived as the search for causes, e. g. [50,87]. However, several explanations are not essentially based on simple causal relations but on other approaches such as identification, models, analogies, formal linguistics, laws of association, laws of co-existence, variational principles, among others [83]. Berger [7] underlines the prominence of this question when attempting to characterize the explanations provided by nonlinear dynamical modeling:

> Mathematical modeling [is recognized] as one of the central activities of science, and it is reasonable to say that modeling explanations dramatically increase our understanding of the world. But the modeling explanations found in contemporary scientific research show that the interesting claims of causal accounts are untenable … An adequate account of scientific explanation must accommodate modeling explanations, because they are simply too central to ignore (pp. 329–330).

The main goal of this section is to explore the position and the characterization of the kind of explanation pursued in system dynamics. This characterization fits with the presented core of SD and the presentationalist stance. And again various clarifications will be needed along the way.

Causality

The difficult issue of causality can be treated in several senses. In the first place, a possible relationship associated with the term "determinism" on human behavior is dismissed with a previous argument: SD causal models does not point at supposed laws of causality governing human action [55,56]. What is more interesting is to investigate the concept of causality as such in SD models; after all, a large part of SD modeling relies on what are known as "causal"-loop diagrams. Forrester emphasized the term *interrelationships* [24] where feedback loops are understood as closed informational paths connecting in a sequence decisions that control actions [25]; he labeled it as a "circular cause-and-effect structure" (p. 1–9). In fact the development of *causal*-loop diagrams has become important in SD practice; in particular, flow diagrams were initially recognized as useful pictorial representations that help to formulate and communicate the structure of a dynamic model [24]. These models are ultimately theories of behavior, surely causal theories of behavior. But consistent with presentationalism, these theories are sourced in the mental models of the modeler, there is no direct connection to an alleged causation in a real world. Sterman summarizes a definition: "a causal diagram consists of variables connected by arrows denoting the causal influences among the variables" [98]; here, every link represents *what the modeler believes is* a causal relationship between two variables; this causal attribution is seen as a central feature of mental models, as Sterman also stresses "we all create and update cognitive maps of causal connections among entities and actors … Within a causal field, people use various cues to causality including temporal and spatial proximity of cause and effect, temporal precedence of causes, covariation, and similarity of cause and effect" [98]. Again, the core of the discussion should be driven by the concept of mental model in order to deliver a clear discussion on this view of causality held in SD. This section outlines a framework.

Most of the time we seem to hold a strong causal view of the world. In particular, the causal relationship – whatever that could be – tends to be the source of explanatory power, i. e. the *explanans*, and one usual source of validity, that is, *for having a relevant valid explanation we must have a causal relationship*. That is the usual principle, and it is usually associated to the term "determinism". Recalling the discussion on presentationalism, one should note that – as Hesslow [42] underlines – if we are going to retain a Humean view of the world, i. e. consistent with idealism, then it seems that we have two different paths, probabilistic or deterministic. The latter one is of interest here (for the probability account of causation see e. g. [43]). Within a deterministic approach, a cause is always sufficient condition or a part of a sufficient condition for the effect, that is, if A_t is a nonsufficient cause of $B_{t'}$, then there must be some auxiliary condition $C_{t''}$, [with $t'' < t'$], such that A_t in combination with $C_{t''}$ is sufficient for $B_{t'}$. Hesslow clarifies this "sufficiency principle" [42]:

> The deterministic approach has been something of a received view of causation. This view, which we may call the 'sufficiency principle' is also common among scientists. The sufficiency principle is not in itself strictly deterministic. It does not mean that every event has a sufficient cause, only that if an event has a cause, then it has a sufficient one. However, it seems that the popularity of the sufficiency principle is a reflection of a widely spread, though usually implicit, commitment to the stronger thesis, that every event has a sufficient cause (p. 592).

The cited paper of Hesslow aims to show that the deterministic approach is superior to the probabilistic one, that is, the idea that the probabilistic account presupposes determinism. Indeed what is happening is the trap of Hume – so to speak. The issue in hands is illustrated with the *Humean fork*: based on observation of constant conjunction of events – altogether with temporal priority, i. e. the cause is observed prior to its effect – we *suppose* a causal connection between them – see summaries in e. g. [46,70] and the original work of Hume [45]. These *suppositions* are arranged in our mental models, that is, in our theories about what we assume as the relevant causal connections that we suppose so as to explain the world.

With this framework in mind, the network of causal connections that the SD modeler believes to be relevant indicates a sort of "sufficiency principle", but of a special kind. It is sufficient relative to the purpose of the model, as it has been indicated above. And it is sourced in the mental model since SD is seen as a vehicle for learning and not as a device for operating a "real world". How are these causal relationships portrayed? In a generic form, a feedback loop based stance is based on the fact that decisions in a time t affect the environment which affects again the future assessment of the situation in a time $t'(t' > t)$ which usually is the base for new upcoming decisions and actions taken in a time t'' $(t'' > t')$ and so on. These relationships are not necessarily close in space or time. Furthermore, a related issue is what can be labeled as "multiple causality". The complexity of social systems is a current concern for social scientists; the multiple interactions among several agents, actors, or entities, is what has been distinguished as the key to study and to understand complex systems because of the recognition of our inability to deal with them based on traditional incomplete simple-causality thinking – the assumption that explanation of phenomena can be satisfactory or even sufficient based in simple unidirectional causal relationships between variables or constructs. What is more, feedback loops have important implications associated with counterintuitive behavior that usually we do not consider easily or that we misunderstand; they have a key role in complex settings; in fact they are, to a large extent, responsible for the arising of complex behaviors; this is a central affirmation in SD [26,98]. The simplest example might be the tendency that we have to infer linear growth from single first order positive feedback loops. But the behavior here actually is exponential. Let S be the state of a system and g the constant fractional growth rate, the linear first order positive loop and its solution are:

$$\frac{\mathrm{d}S}{\mathrm{d}T} = gS$$

and it has as solution:

$$S = S_0 \mathrm{e}^{gt} .$$

The central question is: are these causal theories, portrayed in system dynamics, claims about the world? i. e. Are these models assumed true descriptions of the world? Clearly no. From the presentationalist stance of SD, the causal-loop diagram is essentially what the modeler *believes* is the relevant causal network for the problem in hands. It constitutes his theory about it. The source of knowledge is the mental model and causality is only a supposition of the modeler and a way to arrange knowledge, it is not an affirmation of truth about a supposed causal world. And furthermore, causation as such is not the source of explanation provided by SD. In order to clarify this we should take a look to the notion of explanation held in system dynamics.

Mechanism

System dynamics aims to answer *why* questions. This is done generally via the development of *dynamic hypotheses*. A core premise of SD has always been to enhance learning and to provide understanding [30]. This aim has been stated from the very beginning; for instance Forrester asserted in *Industrial Dynamics*: "Our objective is to enhance understanding and to clarify our thinking about the system" p. 57 in [24]).

How is this goal pursued? A SD model should be able to account for a specific problematic behavior which is explained in terms of the structure of the model; here the term "structure" refers to the stock and flow organization, the feedback loops and the rules of interaction [98]. This approach to explanation is known as a dynamic hypothesis and is the core concept in order to provide understanding from a system dynamics point of view; its endogenous character is the chief feature that makes it intelligible; for instance: "One key task in this search for insightful, system level understanding is the telling of 'system stories' – coherent, dynamically correct explanations of how influential pieces of system structure give rise to important patterns of system behavior" (p. 1 in [65]). In fact this task represents one of the more significant research lines [79].

How can we characterize this particular kind of explanation? As a first point, the notion of organized social complexity helps to drive this discussion. A quote borrowed from Hayek illustrates it:

> Where we have to deal with such social wholes we cannot, as we do in the natural sciences, start from the observation of a number of instances which we recognize spontaneously by their common sense attributes as instances of 'societies' or 'economies' ... What we group together as instances of the same collective or whole are different complexes of individual events, in themselves perhaps quite dissimilar, but believed by us to be related to each other in a similar manner: they are classifications or selections of certain elements of a complex picture on the basis of a theory about their coherence (p. 43 in [40]).

Based on the quote above of Hayek, he suggests conceiving the *explanation* as modeling [104], and for social sciences he rejects the usual prediction and control aspirations and asks the reader to focus more on models to explain typical processes [64], he depicts it with biology: "It deals with pattern-building forces, the knowledge of which is useful for creating conditions favorable to the production of certain kinds of results, while it will only in comparatively few cases be possible to control all the relevant circumstances" (as cited in p. 202 in [104]). Hayek calls it "explanation of the principle". Essentially he means the explanation of a *kind* of phenomena instead of particular events. As another example consider mathematics: "A set of equations which shows merely the form of a system of relationships but does not give the values of the constants contained in it, is perhaps the best general illustration of an explanation merely of the principle on which any phenomenon is produced" (p. 291 in [39]). This analogy illustrates the notion of abstract relations that would build an "explanation of the principle" which can be associated with *mechanism* as the source of explanatory power – instead of causality.

Again, a clarification is needed given the widespread identification of the term "mechanism" with ontic commitments. Fundamentally mechanism is a kind of *explanation*. It should be noticed that "mechanism" refers to *epistemological* issues. However, the term is habitually associated with assumptions about reality. But Hogben already clarified in 1930:

> In any discussion between the two [mechanist and holist or vitalist], the combatants are generally at cross purposes. The mechanist is primarily concerned with an epistemological issue. His critic has always an ontological axe to grind. The mechanist is concerned with how to proceed to a construction which will represent as much about the universe as human beings with their limited range of receptor organs can agree to accept. The vitalist or holist has an incorrigible urge to get behind the limitations of our receptor organs and discover what the universe is really like (1930, p. 100, as cited in [12], p. 347).

The explanatory notion of mechanism is well underlined by Grene: "Let us look for a mechanism which might underlie the phenomena we hope to understand, seeking wherever we may relevant sources from which to derive ... an analogue of a possible mechanism ... [Such an explanation is of value because it tells us] *how in fact those phenomena are produced*" (as cited in [12], p. 346, emphasis original).

However, there is still no agreement about what a mechanism is and how it appears to succeed in science as a way to provide understanding. Perhaps the most complete account is the one of Tabery [99] who proposes integrating two complementary points of view. These two aspects are (i) the interactions among several parts and, (ii) the activities associated with these interactions. Both characteristic are taken as necessary for having a mechanism-based explanation.

On the one hand there is the emphasis on interactions, a thesis supported by Glennan [36] with a central concern on the nature of *complex systems* since the role of parts and its interactions are conceived as indispensable. The work of Bechtel and Richardson [5] develops the association of mechanism and complex systems (in biology and psychology) and emphasizes the tasks of decomposition and localization as the heuristics in order to uncover mechanisms. This position replies to the conventional view of a mechanism as merely the interactions between causal processes as the essential *explanans*; Glennan [36] stresses: "A mechanism for a behavior is a complex system that produces that behavior by the interaction of a number of parts, where the interactions between parts can be characterized by direct, invariant, change-relating generalizations" (p. 344). Glennan clarifies also that he avoids using the term "causal-law" and instead uses "change-relating generalizations" because these relations are not exception-less as traditionally a law is understood. In addition, he emphasizes the very different character of such account which is in opposition to the traditional causal view in which mechanisms are sequences or chains of events leading up to a particular event – which is often associated in systems theory literature with "linear thinking".

On the other hand we have activities. Machamer, Darden and Craver [62] emphasize this aspect adding that mechanisms are not only inter-connected entities but also activities producing regular changes from initial to finish conditions; they call themselves dualists since for them both notions – entities and activities – are necessary to constitute a mechanism: "The organization of these entities and activities determines the ways in which they produce the phenomenon. Entities often must be appropriately located, structured, and oriented, and the activities in which they engage must have a temporal order, rate, and duration" (p. 3). It is important to underline their critique to Glennan's view arguing that the concept of *activity* is fundamental to understand the changes produced (because of the activities) through the process and not only as the black-box view of change of states or change of properties of the inter-connected entities; they picture it clearly with the following statement: "it is not the penicillin that causes the pneumonia to disappear, but what the penicillin does" (p. 6). Furthermore, in order to account for a mechanism they emphasize three distinctions: set-up conditions (as *part* of the mechanism, not as a sort of input; this includes relevant entities and their properties and initial states), intermediate activities (including also relevant entities, properties, and an intelligible account of the activities that link them) and termination conditions (such as privileged endpoints, equilibrium states or the final stage of some unitary integral process). They also draw attention to the fact that mechanisms take place in nested multi-level hierarchies and that they usually are not full pictures but truncated abstract accounts – *a mechanism schema* – depending on the required level of detail or aggregation.

System dynamics modeling is perhaps one of the best ways to picture this kind of explanation. One can distinguish two main components in the structure of SD models: the physical and institutional assumptions – including the chosen parts/variables and the interconnections between them, and the decision rules of the agents [98]. The interplay between the physical structure and the associated decision rules as the explanation for behavior is a foundational aspect. Indeed the interconnections and the activities needed to account for a mechanism are included in these system dynamic models structures. Specifically, the activities producing change can be referenced in the links of the models and the decision rules describe how the interactions produce certain activities. The whole set of initial and final conditions and the inter-connected parts engaged in producing activities characterize a mechanism. For example, the simplest mechanism is perhaps a single feedback loop. The set-up conditions are the initial values of the variables involved, the termination condition is the endpoint of the loop which can be accounted in a mechanism as "the final stage of what is identified as a unitary, integral process" (p. 12 in [62]). The intermediate activities are depicted by the links and the application of the decision rules. For instance, a simple positive first order loop can produce exponential behavior beyond the particular values of the variables involved. More intricate structures are the source of different behaviors, i. e. the change in the values or patterns of variables through time.

This stance fits an explanation of an abstract principle in the sense of Hayek. The structure is the source of explanation of patterns of behavior, i. e. the change in the values or patterns of variables of interest through time. This is known in SD as a dynamic hypothesis [98]. With this focus on aggregate patterns – instead of individual events – as consequence of the structure, it can be said that the "causal" mechanism is indeed the loop structure of the system, or the particular and relevant feedback substructures of the model that may explain the behavior; for instance Richardson [78] illustrates it in this way: "The 'cause' of an arms race is viewed not as a given event or even a given sequence of events, but as a feedback structure dominated by self-reinforcing positive loops, within which events take place" (p. 338). These types of explanations are based on *mechanisms* as explanatory power and not in simple causal relationships as the source of explanation, even less in (substantivalist) causality, i. e. change in singular properties/entities. Furthermore, these hypotheses are developed for each problem consistent with the mental models of the modelers. This is why system dynamics is not committed to specific theories and only to the explanation of problematic behaviors in terms of structure of the model in order to enhance learning and decision-making.

A particular remark must be made. It can be noticed a natural link between mechanism and the idea of "generic structure". This latter expression has been used in different senses in SD literature. Lane and Smart [58] trace the evolution of this concept – see also [71], they identify three different interpretations. One of these view cannot be connected with mechanism, the one popularized by Senge, e. g. [92,109,110], usually known under the expression "system archetypes". The lack of computer simulation within this interpretation points at a problem of validity in its scope and claims, as Lane and Smart discuss [58], e. g. since this perspective skips the possibility of formal computer model building then the relation between structure and behavior is weak and the mentioned approach becomes just a hasty shortcut from problematic behavior to insights and principles, and without the experimental spirit of SD for enhancing learning. But two other notions are relevant. On the one hand, generic structures can be conceived as general models (theories of behavior) of a class of systems that are associated with a domain of application, e. g. urban development, supply chain, economic growth. Lane and Smart label these structures as "canonical situation models". On the other hand, a generic structure may refer to theories of mathematical structures (feedback loops, levels, rate equations, etc.) that generate corresponding dynamic behaviors, i. e. "systems belong to the same class if they can be represented by the same structure … This dynamic structure when abstracted from any application domain data defines the class of system" (p. 93 in [58]); therefore they offer transferability of structure across diverse domains. These models can be labeled as "abstracted micro-structures", e. g. patterns of exponential growth, goal seeking, oscillation, etc. [58,98]. Both interpretations look for establishing a general class of models that formally link structure with behavior and they constitute an important line of research. These developments contribute to different aspects of SD practice; in particular they directly fuel the processes of conceptualization and formal model construction (e. g. see [31,58,98]), and more important, they enhance understanding and the improvement of our mental models as long as we can exploit the powerful idea of having general classes of models, either within a domain of application or across different domains by transferring structures across them.

Consequently, system dynamics explanations can be characterized as *mechanisms*, since there can be found the source of explanatory power. In spite of its causal diagrams, the *explanans*, i. e. that which does the explaining, is based on mechanisms – dynamic hypotheses based on structures; and the problematic behavior is the *explanandum*, i. e. that which is explained. Following Glymour: "Remains, however, a considerable bit of science that sounds very much like explaining, and which perhaps has causal implications, but which does not seem to derive its point, its force, or its interest from the fact that it has something to do with causal relations (or their absence)" (as cited in [p. 212 in 83]). The theories built with SD are essentially structure-based and not content-based (substantivalist) explanations i. e. they are not associated with intrinsic properties of objects or entities but with the consequences of processes and activities entrenched in relevant parts of the structure of the model. Recalling Hayek who identifies explanation with modeling, it is interesting to notice the range of his thought expressed half a century ago and that accurately illustrates SD explanations:

> Any model defines a certain range of phenomena which can be produced by the type of situation which it represents. We may not be able directly to confirm that the causal mechanism determining the phenomenon in question is the same as that of the model. But we know that, if the mechanism is the same, the observed structures must be capable of showing some kinds of action and unable to show others (p. 221 in [38]).

A further reminder follows. Since SD was previously identified with instrumentalism, then a mechanism is not to be taken as a description; here a mechanism is an instrument for arranging observations. It should be kept in mind that mechanism is a kind of explanation which refers to epistemological issues.

The popularity of simple causality as the way to characterize the explanation of phenomena contrasts with the assumptions made in SD: structures that generate processes responsible for behavior. This is consistent with the purpose of system dynamics simulation which might be oriented to activities such as theoretical-representations building, articulation and testing in order to learn in and about complex systems [98]. System dynamics uses simulation as a method which is different from the traditional inductive logic of research that deals with single instances which attempts to confirm theories via repeated observation. However, SD does not dismiss presentationalism as it was shown. This should be highlighted as an important and distinctive characteristic of SD. Though there is a commitment with a real world, justification is rooted in social processes and on mental models, and it is also relative to the purpose of the model. Furthermore, the goal is to enhance our decision-making processes by improving our mental models. How can we characterize such method? A short comment follows.

Simulation and Method

In a plain sense "simulation means driving a model of a system with suitable inputs and observing the corresponding outputs" (p. 23 in [2]). But simulation actually is not just a matter of number crunching. Its scope is broader. And it represents another challenge for philosophy of science. Winsberg [107] illustrates it:

> Typically, to a philosopher of science, epistemological issues arise when we try to justify high level the-

> oretical claims based on low level data or specific observational reports. But simulation is about starting with theory and working your way down. This kind of epistemology is, to the philosopher of science, a curious beast. It is an epistemology that is concerned with justifying inferences from a theory to its application – an inference that most philosophy of science has assumed is deductive and consequently not in need of justification (p. S447).

What is simulation in SD? It can be affirmed that it is a technique able to represent and test theoretical concepts and not only – in a narrow sense – a tool to just solve mathematical problems. Besides, the emphasis on processes, on patterns of collective action and on the relations between components and its dynamic consequences can be better addressed with simulation because of its capacities to represent these issues with fewer restrictions than other approaches [35]. But there is more. Simulation reflects a very different attitude. This way of inquiry suggests a whole different and new scientific methodology [82,106,107]. Winsberg emphasizes that "simulation represents an entirely new mode of scientific activity – one that lies between theory and experiment ... a form of theory articulation or 'model building' (pp. 117, 119 in [106]). Axelrod [2] indeed suggests "a third way" of doing science:

> Simulation as a way of doing science can be contrasted with the two standard methods of induction and deduction ... Simulation is a third way of doing science. Like deduction, it starts with a set of explicit assumptions. But unlike deduction, it does not prove theorems. Instead, a simulation generates data that can be analyzed inductively. Unlike typical induction, however, the simulated data comes from a rigorously specified set of rules rather than direct measurement of the real world (p. 24).

Moreover, its strength rests on the capacity for conducting *experiments* [82]. This emphasis on experimentation is the key to understand why this approach is different. Our mental models are nothing more than theoretical models that attribute properties and relations to the systems they represent; the relevance of these theoretical models depends on the purpose of the model. And computer simulation simply permits experiments of these (theoretical) models. This is where the novelty and the power of this methodology are to be found, in the very iterative process of model building and experimentation via simulation. This position contrasts with the traditional prominence of assumed representational capacities of *theories* and *models* where usually the emphasis has been placed, see [107]. But computer simulation has a distinct epistemology [105] that emphasizes *the process of modeling*. The method was demarcated by Forrester [24] in *Industrial Dynamics*:

> Simulation consists of tracing through, step by step, the actual flows of orders, goods, and information, and observing the series of new decisions that take place ... This is the counterpart of trying a new policy or organizational structure in the real system... After a simulation run comes interpretation of the results. Did it turn out as expected? If not, why? As the experiment is examined, new questions arise ... This is a process of invention and trial ... Each simulation result teaches, and it also prompts additional questions ... Such experimentation will yield new insights into the characteristics of the system that the model represents (pp. 23, 44–45, 55).

Hence, the method of simulation through continued experimentation is aimed at providing better understanding of the modeled system. As it was mentioned, the method calls attention to the *process* – see also [44,89]. Indeed SD aims at developing a *modeling* culture (consistent with [80]) that gives emphasis to model building as an ongoing dialectic between stakeholders instead of a mapping exercise concerned on the efficacy of the model itself.

Why is fundamental the use of the computer? Perhaps the best answer is provided again by Forrester [26]:

> We stress the importance of being explicit about assumptions and interrelating them in a computer model ... The most important difference between the properly conceived computer model and the mental model is in the ability to determine the dynamic consequences when the assumptions within the model interact with one another. The human mind is not adapted to sensing correctly the consequences of a mental model ... The computer model ... is a statement of system structure. It contains the assumptions being made about the system ... Generally, the consequences are unexpected (pp. 213–215).

The shortcomings of our mental models coupled with the complexity of the systems we model lead to the use of the computer. Sterman [101] summarizes these drawback with aspects such our flawed cognitive maps and our erroneous inferences about dynamics. As it was shown, the strength of explanation and understanding in SD is not in the causal models as such; the heart lies in the development of dynamic hypotheses with the use of simulation in order to enhance our understanding and our decision-making processes. Explanations are posed under the notion of mechanism; but this is an iterative process that seeks a central aim: to improve our own theories about the world.

Future Directions

There are several aspects to develop based on this reflection on various assumptions behind system dynamics. The central argument was built around the position known as presentationalism. This stance integrates and informs many of the debates on related subjects, e. g. validation, and it characterizes the initial purposes and assumptions held in the field. However, there are a number of lines to emphasize and to develop.

We have focused on the role of the idea of "mental model" for the practice of SD. Yet, it can be seen utilization of SD models under assumptions which are closer to a naive realism that seem to ignore the purposes of enhancing understanding and learning processes. This article should help to underscore that system dynamics is less naive – and hence more powerful – when we recognize a presentationalist stance which means that our theories about the world are just that, *theories* based on our experience and on the operations of our own mind. We can improve these theories with the use of system dynamics, that is, making our assumptions about systems explicit and using simulation as a method for enhancing understanding and for developing explanations that guide our processes of systems design. This recognition includes the relative nature of justification of knowledge held in SD and the emphasis on mechanism as a powerful way to develop explanations about complex systems.

It has been introduced mechanism as the way to characterize the particular type of explanation pursued in SD. The explanatory force does not rest on causal relations as such but on the structures – physical and decision rules aspects – and on the dynamic processes and activities that explain change. The idea of mechanism is shaped in SD under the expression "generic structure". The focus on understanding behavior in terms of abstract structures is a central line of inquiry. This article underlines the importance of developing such a line of research since it has been located at the core of the kind of knowledge that SD provides. The issue of unification to provide understanding of diverse phenomena is a definitive step in the way to assert that the field progresses as long as broader range of phenomena may be explained with the same mechanism. Should be the advance of system dynamics assessed by the progress in this type of study? Behind this discussion there are major and provocative issues that arise to be developed.

The long debate of qualitative and quantitative modeling is informed by this characterization. It is clear that the issue of explanation compels theorists and practitioners to ask themselves what is the kind of explanation they are pursuing and if such explanations are enough and satisfying; it is worthy to ask for qualitative modeling what kind of understanding it gives and how it can be characterized, in other words, to give an account of *explanans* and *explanandum*. Another line to develop might be oriented around the following question: what would be *essential* criteria for comparing different arranges or modes of organization in order to identifying them as belonging to the same type of mechanism? The identification and search of mechanisms becomes a powerful heuristic for guiding the modeling process.

There are promising suggestions for philosophy of science as well. SD offers guidelines, e. g. can the ways in which system dynamicists work provide meaningful insights, or even concrete accounts, for the philosophical unresolved issue of explanation? The mechanism depicted in system dynamics proposes a kind of explanation that goes beyond the received view based on causation. A related question is whether explanation must always follow a deductive path; the classic models of Hempel, e. g. [41], emphasized the condition of deduction and general laws for having an explanation; however, the explanation in SD is not framed under a deductive schema from universal covering laws, instead it can be conceived as a sort of abductive reasoning based on the understanding of the dynamics of the model as a way to understand the actual behavior it accounts for. A further issue is that in spite of the lack of universality, i. e. no universal laws, SD models aim to provide understanding for a diverse range of phenomena that might share relevant influential structures and similar associated behaviors, that is, it accounts for regularities in order to unify them in a certain kind of explanation under the same explanation. This situation insinuates a flavor of paradox because of the traditional rigid association of universal laws with the explanation of regularities, but in SD there are no general laws though the aim is to explain general regularities. The study of *models* and computer simulation – instead of abstract theories and traditional methodologies – is an additional indication that SD suggests for philosophy of science, including the emphasis on the modeling process.

Bibliography

1. Aronson JL (1988) Testing for Convergent Realism. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol One: Contributed Papers. University of Chicago Press, Chicago, pp 188–193

2. Axelrod R (1997) Advancing the art of simulation in the social science. In: Conte R, Hegselmann R, Terna P (eds) Simulating Social Phenomena. Springer, Berlin, pp 21–40
3. Barlas Y, Carpenter S (1990) Philosophical roots of model validation: two paradigms. Syst Dyn Rev 6:148–166
4. Bartley III WW (1987) Philosophy of biology versus philosophy of physics. In: Radnitzky G, Bartley III WW (eds) Evolutionary epistemology, rationality, and the sociology of knowledge. Open Court, La Salle, pp 7–45
5. Bechtel W, Richardson RC (1993) Discovering complexity: Decomposition and localization as strategies in scientific research. Princeton University Press, Princeton
6. Benjamin AC (1941) Modes of scientific explanation. Philos Sci 8:486–492
7. Berger R (1998) Understanding science: Why causes are not enough. Philos Sci 65:306–332
8. Berkeley G (1948–1957) The works of George Berkeley, Bishop of Cloyne. Thomas Nelson and Sons, London
9. Black M (1934) The principle of verifiability. Analysis 2:1–6
10. Blackmore J (1979) On the inverted use of the terms 'Realism' and 'Idealism' among scientists and historians of science. Br J Philos Sci 30:125–134
11. Bowman AA (1916) Kant's phenomenalism in its relation to subsequent metaphysics. Mind, New Ser 25:461–489
12. Brandon RN (1984) Grene on mechanism and reductionism: more than just a side issue. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol II: Symposia and Invited Papers. University of Chicago Press, Chicago, pp 345–353
13. Brown HI (1990) Cherniak on scientific realism. Br J Philos Sci 41:415–427
14. Burrell G, Morgan G (1979) Sociological paradigms and organizational analysis. Heinemann, London
15. Chapin JP (1941) Idealism and its relation to science. Philosophy of Science 8:142–146
16. Checkland P, Pidd M, Morecroft J (2004) Working ideas, insights for systems modeling: The Broader community of systems thinkers. In: Kennedy M, Winch G, Langer R, Rowe J, Yanni J (eds) Proceedings of the 22nd International Conference of the System Dynamics Society, Keble College, University of Oxford, England. System Dynamics Society, Albany
17. Coyle G (2000) Qualitative and quantitative modelling in system dynamics: some research questions. Syst Dyn Rev 16:225–244
18. Coyle RG (1979) Management system dynamics. Wiley, Chichester
19. Deetz S (1996) Describing differences in approaches to organization science: Rethinking Burrell and Morgan and their legacy. Organ Sci 7:191–207
20. Dickson M (1995) An empirical reply to empiricism: Protective measurements opens the door for quantum realism. Philos Sci 62:122–140
21. Doyle JK, Ford DN (1998) Mental models concepts for system dynamics research. Syst Dyn Rev 14:3–29
22. Doyle JK, Ford DN (1999) Mental models concepts revisited: some clarifications and a reply to Lane. Syst Dyn Rev 15:411–415
23. Flood R, Jackson M (1991) Creative problem solving. Wiley, Chichester
24. Forrester JW (1961) Industrial Dynamics. Press MIT, Cambridge
25. Forrester JW (1971) Principles of Systems. Wright-Allen Press, Cambridge
26. Forrester JW (1975) Counterintuitive behavior of social systems. In: Collected papers of Jay W. Forrester. Wright-Allen Press, Cambridge, pp 211–244
27. Forrester JW (1975) Industrial Dynamics: A Major breakthrough for decision makers. In: Collected papers of Jay W Forrester. Wright-Allen Press, Cambridge, pp 1–29
28. Forrester JW (1975) Industrial Dynamics – After the first decade. In: Collected papers of Jay W. Forrester. Wright-Allen Press, Cambridge, pp 133–150
29. Forrester JW (1985) "The" model versus a modeling "process". Syst Dyn Rev 1:133–134
30. Forrester JW (1987) Lessons from system dynamics modeling. Syst Dyn Rev 3:136–149
31. Forrester JW (1994) System dynamics, systems thinking, and soft OR. Syst Dyn Rev 10:245–256
32. Forrester JW (2003) Dynamic models of economic systems and industrial organizations. Syst Dyn Rev 19:331–345
33. Fuller S (1994) Retrieving the point of the realism-instrumentalism debate: Mach vs. Planck on science education policy. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol One: Contributed Papers. University of Chicago Press, Chicago, pp 200–208
34. Gershenson C, Aerts D, Edmonds B (2007) Worldviews, science and us. Philosophy and complexity. World Scientific, Singapore
35. Gilbert N, Troitzsch KG (1999) Simulation for the social scientist. Open University Press, Buckingham
36. Glennan SS (2002) Rethinking mechanistic explanation. Philos Sci 69:342–353
37. Guyer P (1983) Kant's intentions in the refutation of idealism. Philos Rev 92:329–383
38. Hayek FA (1955) Degrees of explanation. Br J Philos Sci 6:209–225
39. Hayek FA (1942) Scientism and the study of society, Part I. Economica, New Ser 9:267–291
40. Hayek FA (1943) Scientism and the study of society, Part II. Economica, New Ser 10:34–63
41. Hempel CG, Oppenheim P (1948) Studies in the logic of explanation. Philos Sci 15:135–175
42. Hesslow G (1981) Causality and determinism. Philos Sci 48:591–605
43. Hitchcock CR, Salmon WC (2000) Statistical explanation. In: Newton-Smith WH (ed) A Companion to the philosophy of science. Blackwell Publishers, Malden, pp 470–479
44. Homer JB (1996) Why we iterate: scientific modeling in theory and practice. Syst Dyn Rev 12:1–19
45. Hume D (1740) A treatise of human nature. Oxford University Press, Oxford
46. Humphreys P (2000) Causation. In: Newton-Smith WH (ed) A companion to the philosophy of science. Blackwell Publishers, Malden, pp 31–40
47. Hunter B (1992) Empiricism. In: Dancy J, Sosa E (eds) A Companion to Epistemology. Blackwell Publishers, Oxford, pp 110–115
48. Jackson M (1991) Systems methodology for the management sciences. Plenum Press, New York
49. Jackson M (2003) Systems thinking: Creative holism for managers. Wiley, Chichester

50. Jobe EK (1985) Explanation, causality, and counterfactuals. Philos Sci 52:357–389
51. Kleindorfer GB, Ganeshan R (1993) The philosophy of science and validation in simulation. In: Evans GW, Mollaghasemi M, Russell EC, Biles WE (eds) Proceedings of the 1993 Winter Simulation Conference. IEEE, Piscataway, pp 50–57
52. Kukla A (1995) Scientific realism and theoretical unification. Analysis 55:230–238
53. Lane D (1994) With a little help from our friends: How system dynamics and soft OR can learn from each other. Syst Dyn Rev 10:101–134
54. Lane D (1999) Friendly amendment: A commentary on Doyle and Ford's proposed re-definition of 'mental model'. Syst Dyn Rev 15:185–194
55. Lane D (2000) Should system dynamics be described as a 'Hard' or 'Deterministic' systems approach? Syst Res Behav Sci 17:3–22
56. Lane D (2001) Rerum cognoscere causas: Part I – How do the ideas of system dynamics relate to traditional social theories and the voluntarism/determinism debate? Syst Dyn Rev 17:97–118
57. Lane D (2001) Rerum cognoscere causas: Part II – Opportunities generated by the agency/structure debate and suggestions for clarifying the social theoretic position of system dynamics. Syst Dyn Rev 17:293–309
58. Lane D, Smart C (1996) Reinterpreting 'generic structure': evolution, application and limitations of a concept. Syst Dyn Rev 12:87–120
59. Leplin J (1992) Realism and Methodological Change. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol Two: Symposia and Invited Papers. University of Chicago Press, Chicago, pp 435–445
60. Leplin J (2000) Realism and Instrumentalism. In: Newton-Smith WH (ed) A Companion to the Philosophy of Science. Blackwell Publishers, Malden, pp 393–401
61. Luna-Reyes LF, Andersen DL (2003) Collecting and analyzing qualitative data for system dynamics: methods and models. Syst Dyn Rev 19:271–296
62. Machamer P, Darden L, Craver CF (2000) Thinking about mechanisms. Philos Sci 67:1–25
63. Meadows DH (1980) The Unavoidable A Priori. In: Randers J (ed) Elements of the system dynamics method. Productivity Press, Cambridge, pp 23–57
64. Milford K (1994) In pursuit of rationality. A note on Hayek's The Counter-Revolution of Science. In: Birner J, van Zijp R (eds) Hayek, Co-ordination and Evolution. Routledge, London, pp 323–340
65. Mojtahedzadeh M, Andersen D, Richardson GP (2004) Using digest to implement the pathway participation method for detecting influential system structure. Syst Dyn Rev 20:1–20
66. Morrison M (1988) Reduction and realism. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol One: Contributed Papers. University of Chicago Press, Chicago, pp 286–293
67. Morrison M (1990) Unification, realism and inference. Br J Philos Sci 41:305–332
68. Newton-Smith WH (1988) Modest realism. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol Two: Symposia and Invited Papers. University of Chicago Press, Chicago, pp 179–189
69. Newton-Smith WH (2000) Explanation. In: Newton-Smith WH (ed) A companion to the philosophy of science. Blackwell Publishers, Malden/Oxford, pp 127–133
70. Newton-Smith WH (2000) Hume. In: Newton-Smith WH (ed) A companion to the philosophy of science. Blackwell Publishers, Malden, pp 165–168
71. Paich M (1985) Generic structures. Syst Dyn Rev 1:126–132
72. Pettit P (1992) Realism. In: Dancy J, Sosa E (eds) A companion to epistemology. Blackwell Publishers, Oxford, pp 420–424
73. Popper K (1963) Conjectures and refutations. The growth of scientific knowledge. Routledge and Kegan Paul, London
74. Popper K (1968) The Logic of Scientific Discovery. Hutchinson, London
75. Psillos S (1996) Scientific realism and the "pessimistic induction". Proceedings of the Biennial Meeting of the Philosophy of Science Association, Part I: Contributed Papers. University of Chicago Press, Chicago, pp S306-S314
76. Radnitzky G (1987) In defense of self-applicable critical rationalism. In: Radnitzky G, Bartley III WW (eds) Evolutionary epistemology, rationality, and the sociology of knowledge. Open Court, La Salle, pp 279–312
77. Rescher N (1992) Idealism. In: Dancy J, Sosa E (eds) A companion to epistemology. Blackwell Publishers, Oxford, pp 187–191
78. Richardson GP (1991) Feedback thought in social science and systems theory. Pegasus Communications, Waltham
79. Richardson GP (1996) Problems for the future of system dynamics. Syst Dyn Rev 12:141–157
80. Richardson KA (2002) On the limits of bottom-up computer simulation: Towards a nonlinear modeling culture. In: Sprague RH Jr (ed) Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03). IEEE
81. Roberts EB (1978) System dynamics – An introduction. In: Roberts EB (ed) Managerial applications of system dynamics. Pegasus Communications Inc., Waltham
82. Rohrlich F (1990) Computer simulation in the physical sciences. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol Two: Symposia and Invited Papers. University of Chicago Press, Chicago, pp 507–518
83. Ruben D (1990) Explaining explanation. Routledge, London
84. Runciman WG (1969) What is structuralism? Br J Sociol 20:253–265
85. Russell B (1912) The problems of philosophy. Oxford University Press, Oxford
86. Russell B (1948) Human knowledge: Its scope and limits. Simon and Schuster, New York
87. Salmon W (1992) Explanation. In: Dancy J, Sosa E (eds) A companion to epistemology. Blackwell Publishers, Oxford, pp 129–132
88. Salmon WC (2000) Logical Empiricism. In: Newton-Smith WH (ed) A companion to the philosophy of science. Blackwell Publishers, Malden, pp 233–242
89. Schaffernicht M (2006) Detecting and monitoring change in models. Syst Dyn Rev 22:73–88
90. Schlagel RH (1981) Contextualistic realism. Philos Phenomenol Res 41:437–451
91. Sellars W (1976) Is scientific realism tenable? PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol Two: Symposia and Invited Papers. University of Chicago Press, Chicago, pp 307–334

92. Senge P (1990) The fifth discipline: The art and practice of the learning organization. Doubleday, New York
93. Sidgwick H, Caird E (1880) Kant's refutation of idealism. Mind, New Ser 5:111–115
94. Sinha D (1963) Phenomenology and positivism. Philos Phenomenol Res 23:562–577
95. Smith DW (1982) The realism in perception. Noûs 16:42–55
96. Spender J-C (1996) Making knowledge the basis of a dynamic theory of the firm. Strateg Manag J 17:45–62
97. Stace WT (1944) Positivism. Mind, New Ser 53:215–237
98. Sterman J (2000) Business dynamics. Systems thinking and modeling for a complex world. McGraw-Hill, Boston
99. Tabery JG (2004) Synthesizing activities and interactions in the concept of a mechanism. Philos Sci 71:1–15
100. Taube M (1937) Positivism, science, and history. J Philos 34:205–210
101. Turbayne CM (1955) Kant's refutation of dogmatic idealism. Philos Q 5:225–244
102. Vanderminden P (2006) System dynamics – A field of study, a methodology or both. In: Größler A, Rouwette E, Langer R, Rowe J, Yanni J (eds) 24th International Conference of The System Dynamics Society. Radboud University Nijmegen, System Dynamics Society, Albany
103. Vásquez M, Liz M, Aracil J (1996) Knowledge and reality: some conceptual issues in system dynamics modeling. Syst Dyn Rev 12:21–37
104. Weimer W (1999) Hayek's approach to the problems of complex phenomena: an introduction to the theoretical psychology of The Sensory Order. In: Boettke P (ed) The Legacy of Friedrich von Hayek, vol II. Edward Elgar Publishing Limited, Cheltenham, pp 200–244
105. Winsberg E (1999) Sanctioning models: The epistemology of simulation. Sci Context 12:275–293
106. Winsberg E (2003) Simulated experiments: Methodology for a virtual world. Philos Sci 70:105–125
107. Winsberg E (2001) Simulations, models, and theories: Complex physical systems and their representations. Philos Sci 68:S442–S454
108. Wolstenholme EF (1990) System enquiry. Wiley, Chichester
109. Wolstenholme EF (2003) Towards the definition and use of a core set of archetypal structures in system dynamics. Syst Dyn Rev 19:7–26
110. Wolstenholme EF (2004) Using generic system archetypes to support thinking and modelling. Syst Dyn Rev 20:341–356

System Regulation and Design, Geometric and Algebraic Methods in

ALBERTO ISIDORI[1], LORENZO MARCONI[2]

[1] Department of Informatica e Sistemistica,
University of Rome, La Sapienza, Italy

[2] Center for Research on Complex Automated Systems
(CASY), University of Bologna, Bologna, Italy

Article Outline

Glossary

Exosystem A dynamical system modeling the set of all exogenous inputs (command/disturbances) affecting a controlled plant.

Internal model A model of the exogenous inputs (command/disturbances) affecting a controlled plant, embedded in the interior of the controller.

Generalized tracking problem The problem of designing a controller able to asymptotically track/reject any exogenous command/disturbance in a fixed set of functions.

Observer A device designed to asymptotically track the state of a dynamical system on the basis of measured observations.

Steady state A family of behaviors, in a dynamical system, that are asymptotically approached, as actual time tends to infinity or as initial time tends to minus infinity.

Definition

A central problem in control theory is the design of feedback controllers so as to have certain outputs of a given plant *to track* prescribed reference trajectories. In any realistic scenario, this control goal has to be achieved in spite of a good number of phenomena which would cause the system to behave differently than expected. These phenomena could be endogenous, for instance parameter variations, or exogenous, such as additional undesired inputs affecting the behavior of the plant. In numerous design problems, exogenous inputs are not available for measurement, nor are known ahead of time, but rather can only be seen as unspecified members of a given family of functions. Embedding a suitable "internal model" of such a family in the controller is a design strategy that has proven to be quite successful in handling uncertainties in the controlled plant as well as in the exogenous inputs.

Introduction

The problem of controlling the output of a system so as to achieve asymptotic tracking of prescribed trajectories and/or asymptotic rejection of disturbances is a central problem in control theory. There are essentially three different possibilities to approach the problem: tracking by dynamic inversion, adaptive tracking, tracking via internal models. Tracking by *dynamic inversion* consists in computing a precise initial state and a precise control input (or equivalently a reference trajectory of the state), such that, if the system is accordingly initialized and driven, its output exactly reproduces the reference signal. The computation of such control input, however, requires "perfect knowledge" of the entire trajectory which is to be tracked as well as "perfect knowledge" of the model of the plant to be controlled. Thus, this type of approach is not suitable in the presence of large uncertainties on plant parameters as well as on the reference signal. *Adaptive* tracking consists in tuning the parameters of a control input computed via dynamic inversion in such a way as to guarantee asymptotic convergence to zero of a tracking error. This method can successfully handle parameter uncertainties, but still presupposes the knowledge of the entire trajectory which is to be tracked (to be used in the design of the adaptation algorithm) and therefore an approach of this kind is not suited to the problem of tracking unknown trajectories. Of course, one might consider the problem of tracking a slowly varying reference trajectory as a stabilization problem in the presence of a slowly varying unknown parameter, but this would, in most cases, yield a very conservative solution.

In most cases of practical interest, the trajectory to be tracked (or the disturbance to be rejected) is not available for measurement. Rather, it is only known that this trajectory is simply an (undefined) member in a set of functions, for instance the set of all possible solutions of an ordinary differential equation. These cases include the classical problem of the set point control, the problem of active suppression of harmonic disturbances of unknown amplitude, phase and even frequency, the synchronization of nonlinear oscillations, and similar others. It is in these cases that *tracking via internal models* proves particularly efficient, in its ability to handle simultaneously uncertainties in plant parameters as well as in the trajectory which is to be tracked.

For linear multivariable systems, tracking problems of this kind (those in which the exogenous commands and/or disturbances are only known to be members in the set of solutions of a given ordinary differential equations) have been addressed in very elegant geometric terms by Davison, Francis, Wonham [8,10,11] and others. In particular, one of the most relevant contributions of [11] was a clear delineation of what is known as *internal model principle*, i. e. the fact that the property of perfect tracking is insensitive to plant parameter variations "only if the controller utilizes feedback of the regulated variable, and incorporates in the feedback path a suitably reduplicated model of the dynamic structure of the exogenous signals which the regulator is required to process". Conversely, in a stable-closed loop system, if the controller utilizes feedback of the regulated variable and incorporates an internal model of the exogenous signals, the output regulation property is insensitive to plant parameter variations.

A nonlinear enhancement of this theory, which uses a combination of geometry and nonlinear dynamical systems theory, was initiated in [15,16,19] in the context of solving the problem near an equilibrium, in the presence of exogenous signals which were produced by a Poisson stable system. In particular, in [19] it was shown how the use center manifold theory near an equilibrium determines the necessity of the existence of an internal model whenever one can solve an output regulation problem in spite of (small) parameter uncertainties (see also [3]). Since these pioneering contributions, the theory has experienced a tremendous growth, culminating in the development of design methods able to handle the case of parametric uncertainties affecting the autonomous (linear) system which generates the exogenous signals (such as in [9,23]), the case of nonlinear exogenous systems (such as in [5]), or a combination thereof (as in [22]). The purpose of these notes is to present a self-contained exposition of the fundamentals of these design methods.

The Generalized Tracking Problem

In this article, we address tracking problems that can be cast in the following terms. Consider a finite-dimensional, time-invariant, nonlinear system described by equations of the form

$$\begin{aligned}\dot{x} &= f(w,x,u)\\ e &= h(w,x)\\ y &= k(w,x)\,,\end{aligned} \tag{1}$$

in which $x \in \mathbb{R}^n$ is a vector of state variables, $u \in \mathbb{R}^m$ is a vector of inputs used for *control* purposes, $w \in \mathbb{R}^s$ is a vector of inputs which cannot be controlled and include *exogenous* commands, exogenous disturbances and model uncertainties, $e \in \mathbb{R}^p$ is a vector of *regulated* outputs which include tracking errors and any other variable that needs to be steered to 0, $y \in \mathbb{R}^q$ is a vector of outputs that are available for *measurement* and hence used to feed the device that supplies the control action. The problem is to design a controller, which receives $y(t)$ as input and produces $u(t)$ as output, able to guarantee that, in the resulting closed-loop system, $x(t)$ remains bounded and

$$\lim_{t\to\infty} e(t) = 0\,, \tag{2}$$

regardless of what the exogenous input $w(t)$ actually is.

The ability to successfully address this problem very much depends on how much the controller is allowed to know about the exogenous disturbance $w(t)$. In the ideal situation in which $w(t)$ is available to the controller in real-time, the design problem indeed looks much simpler. This is, though, only an extremely optimistic situation which does not represent, in any circumstance, a realistic scenario. The other extreme situation is the one in which nothing is known about $w(t)$. In this, pessimistic, scenario the best result one could hope for is the fulfillment of some prescribed ultimate bound for $|e(t)|$, but certainly not a sharp goal such as (2). A more comfortable, intermediate, situation is the one in which $w(t)$ is only known *to belong to a fixed family* of functions of time, for instance the family of all solutions obtained from a fixed ordinary differential equation of the form

$$\dot{w} = s(w) \tag{3}$$

as the corresponding initial condition $w(0)$ is allowed to vary on a prescribed set. This situation is in fact sufficiently distant from the ideal but unrealistic case of perfect knowledge of $w(t)$ and from the realistic but conservative case of totally unknown $w(t)$. But, above all, this way of thinking at the exogenous inputs covers a number of cases of major practical relevance. There is, in fact, abundance of design problems in which parameter uncertainties, reference commands and/or exogenous disturbances can be modeled as functions of time that satisfy an ordinary differential equation.

The control law is to be provided by a system modeled by equations of the form

$$\begin{aligned} \dot{\xi} &= \varphi(\xi, y) \\ u &= \gamma(\xi, y) \end{aligned} \tag{4}$$

with state $\xi \in \mathbb{R}^\nu$. The initial conditions $x(0)$ of the *plant* (1), $w(0)$ of the *exosystem* (3) and $\xi(0)$ of the *controller* (4) are allowed to range over fixed *compact* sets $X \subset \mathbb{R}^n$, $W \subset \mathbb{R}^s$ and, respectively, $\Xi \subset \mathbb{R}^\nu$. All maps characterizing the model of the controlled plant, of the exosystem and of the controller are assumed to be sufficiently differentiable.

The problem which will be studied, known as the *generalized tracking problem* (or *problem of output regulation* or also *generalized servomechanism problem*) is to design a feedback controller of the form (4) so as to obtain a closed loop system in which all trajectories are bounded and the regulated output $e(t)$ asymptotically decays to 0 as $t \to \infty$. More precisely, it is required that the composition of (1), (3) and (4), that is the *autonomous* system

$$\begin{aligned} \dot{w} &= s(w) \\ \dot{x} &= f(w, x, \gamma(\xi, k(w, x))) \\ \dot{\xi} &= \varphi(\xi, k(w, x)) \end{aligned} \tag{5}$$

with output

$$e = h(w, x)\,,$$

be such that:

- The positive orbit of $W \times X \times \Xi$ is bounded, i. e. there exists a bounded subset S of $\mathbb{R}^s \times \mathbb{R}^n \times \mathbb{R}^\nu$ such that, for any $(w_0, x_0, \xi_0) \in W \times X \times \Xi$, the integral curve $(w(t), x(t), \xi(t))$ of (5) passing through (w_0, x_0, ξ_0) at time $t = 0$ remains in S for all $t \geq 0$.
- $\lim_{t\to\infty} e(t) = 0$, uniformly in the initial condition, i. e., for every $\varepsilon > 0$ there exists a time $\bar{t}$, depending only on ε and *not on* (w_0, x_0, ξ_0), such that the integral curve $(w(t), x(t), \xi(t))$ of (5) passing through (w_0, x_0, ξ_0) at time $t = 0$ satisfies $\|e(t)\| \leq \varepsilon$ for all $t \geq \bar{t}$.

The Steady-State Behavior of a System

Limit Sets

The generalized tracking problem can be seen as the problem of forcing in the plant, by means of an appropriate control input $u(t)$, a response $x(t)$ that asymptotically compensates the effect, on the regulated variable $e(t)$, of the exogenous input $w(t)$. The classical way in which the problem is addressed for linear, time-invariant systems, when the exosystem is a neutrally stable linear system, is to seek a controller forcing in the associated closed-loop system (5) a (stable) "steady state" behavior entirely contained in the kernel of the map defining the tracking error e. Thus, it is natural to expect that a similar tool should also be effective in the more general setting considered here. It appears, though, that a rigorous investigation of the concept of "steady state", beyond the classical domain of linear system theory, had never been fully pursued.

Motivated by the current practice in linear system theory, the "steady state" behavior of a dynamical system can be viewed as a kind *limit* behavior, approached either as the *actual* time t tends to $+\infty$ or, alternatively, as the *initial* time t_0 tends to $-\infty$. Relevant, in this regard, are certain concepts introduced by G.D. Birkhoff in his classical

1927 essay, where he asserts that “with an arbitrary dynamical system ... there is associated always a closed set of ‘central motions’ which do possess this property of regional recurrence, towards which all other motions of the system in general tend asymptotically” (see p. 190 in [2]). In particular, a fundamental role is played by the concept of ω-limit set of a given point, which is defined as follows. Consider an *autonomous* ordinary differential equation

$$\dot{x} = f(x) \tag{6}$$

with $x \in \mathbb{R}^n$, $t \in \mathbb{R}$. It is well known that, if $f\colon \mathbb{R}^n \to \mathbb{R}^n$ is locally Lipschitz, for any $x_0 \in \mathbb{R}^n$, the solution of (6) with initial condition $x(0) = x_0$, denoted by $x(t, x_0)$, exists on some open interval of the point $t = 0$ and is unique.

Assume, in particular, that $x(t, x_0)$ is defined for all $t \geq 0$. A point x is said to be an ω-limit *point* of the motion $x(t, x_0)$ if there exists a sequence of times $\{t_k\}$, with $\lim_{k\to\infty} t_k = \infty$, such that

$$\lim_{k\to\infty} x(t_k, x_0) = x\,.$$

The ω-limit *set* of a point x_0, denoted $\omega(x_0)$, is *the union* of all ω-limit points of the motion $x(t, x_0)$.

It is obvious from this definition that an ω-limit point *is not* necessarily a limit of $x(t, x_0)$ as $t \to \infty$, as the solution in question may not admit any limit as $t \to \infty$. It happens though, that if the motion $x(t, x_0)$ is *bounded*, then $x(t, x_0)$ asymptotically approaches *the set* $\omega(x_0)$. This property is precisely described in what follows [2].

Lemma 1 *Suppose there is a number M such that $\|x(t, x_0)\| \leq M$ for all $t \geq 0$. Then, $\omega(x_0)$ is a nonempty compact connected set, invariant under (6). Moreover, the distance of $x(t, x_0)$ from $\omega(x_0)$ tends to 0 as $t \to \infty$.*

One of the remarkable features of $\omega(x_0)$, as indicated in this Lemma, is the fact that this set is *invariant* for (6). Invariance means that for any initial condition $\bar{x}_0 \in \omega(x_0)$ the solution $x(t, \bar{x}_0)$ of (6) exists *for all* $t \in (-\infty, +\infty)$ and that $x(t, \bar{x}_0) \in \omega(x_0)$ for all such t. Put in different terms, the set $\omega(x_0)$ is filled by motions of (6) which are bounded backward and forward in time. The other remarkable feature is that $x(t, x_0)$ *approaches* $\omega(x_0)$ as $t \to \infty$, in the sense that the distance *of the point* $x(t, x_0)$ (the value at time t of the solution of (6) starting in x_0 at time $t = 0$) *from the set* $\omega(x_0)$ tends to 0 as $t \to \infty$.

Since any motion $x(t, x_0)$ which is bounded in positive time asymptotically approaches the ω-limit set $\omega(x_0)$ as $t \to \infty$, one may be tempted to look, for a system (6) in which *all* motions are bounded in positive time, at the *union* of the limit sets of all points x_0, i. e. at the set

$$\Omega = \bigcup_{x_0 \in \mathbb{R}^n} \omega(x_0)$$

and to say that the system is in steady state if its state $x(t)$ evolves in the (invariant) set Ω. There is a major drawback, though, in taking this as definition of “steady state” behavior of a nonlinear system: the convergence of $x(t, x_0)$ to Ω is not guaranteed to be *uniform* in x_0, even if the latter ranges over a compact set (see, e. g [7]).

One of the main motivations for looking into the concept of steady state is the aim *to shape* the steady state response of a system to a given (or to a given family of) forcing inputs. But this motivation looses much of its meaning if the time needed to get within an ε-distance from the steady state may grow unbounded as the initial state changes (even when the latter is picked within a fixed *bounded* set). In other words, *uniform* convergence to the steady state (which is automatically guaranteed in the case of linear systems) is an indispensable feature to be required in a nonlinear version of this notion. The set Ω, the union of all ω-limit points of all points in the state space does not have this property of uniform convergence, but there is a larger set which does have this property. This larger set, known as the ω limit set *of a set*, is precisely defined as follows.

Consider again system (6), let B be a subset of $\mathbb{R}^n$ and suppose $x(t, x_0)$ is defined for all $t \geq 0$ and all $x_0 \in B$. The ω-limit set of B, denoted $\omega(B)$, is the set of all points x for which there exists a sequence of pairs $\{x_k, t_k\}$, with $x_k \in B$ and $\lim_{k\to\infty} t_k = \infty$ such that

$$\lim_{k\to\infty} x(t_k, x_k) = x\,.$$

It is clear from the definition that if B consists of only one single point x_0, all x_k’s in the definition above are necessarily equal to x_0 and the definition in question reduces to the definition of ω-limit set of a point, given earlier. It is also clear form this definition that, if for some $x_0 \in B$ the set $\omega(x_0)$ is nonempty, all points of $\omega(x_0)$ are points of $\omega(B)$. In fact, all such points have the property indicated in the definition, if all the x_k’s are taken equal to x_0. Thus, in particular, if all motions with $x_0 \in B$ are bounded in positive time,

$$\bigcup_{x_0 \in B} \omega(x_0) \subset \omega(B)\,.$$

However, the converse inclusion is not true in general.

The relevant properties of the ω-limit set of a set, which extend those presented earlier in Lemma 1, can be summarized as follows [14].

Lemma 2 *Let B be a nonempty bounded subset of $\mathbb{R}^n$ and suppose there is a number M such that $\|x(t, x_0)\| \leq M$ for all $t \geq 0$ and all $x_0 \in B$. Then $\omega(B)$ is a nonempty compact* *set, invariant under (6). Moreover, the distance of $x(t, x_0)$ from $\omega(B)$ tends to 0 as $t \to \infty$, uniformly in $x_0 \in B$. If B is connected, so is $\omega(B)$.*

Thus, as it is the case for the ω-limit set of a point, we see that the ω-limit set of a bounded set, being compact and invariant, is filled with motions which exist for all $t \in (-\infty, +\infty)$ and are bounded backward and forward in time. But, above all, we see that the set in question is *uniformly* approached by motions with initial state $x_0 \in B$, a property that the set Ω does not have. Note also that the set of all such trajectories is a "behavior", in the sense of J.C. Willems [25].

The set $\omega(B)$, as shown in the previous Lemma, asymptotically attracts, as $t \to \infty$, all motions that start in B. Since the convergence to $\omega(B)$ is uniform in x_0, it is also true that, whenever $\omega(B)$ is contained in the interior of B, the set $\omega(B)$ is *asymptotically stable*, in the sense of Lyapunov (see [14]).

The Steady State Behavior of a Nonlinear System

Consider now again system (6), with initial conditions in a closed subset $X \subset \mathbb{R}^n$. Suppose the set X is *positively invariant*, which means that for any initial condition $x_0 \in X$, the solution $x(t, x_0)$ exists for all $t \geq 0$ and $x(t, x_0) \in X$ for all $t \geq 0$. The motions of this system are said to be *ultimately bounded* if there is a bounded subset B with the property that, for every compact subset X_0 of X, there is a time $T > 0$ such that $\|x(t, x_0)\| \in B$ for all $t \geq T$ and all $x_0 \in X_0$. In other words, if the motions of the system are ultimately bounded, every motion eventually enters and remains in the bounded set B.

Suppose the motions of (6) are ultimately bounded and let $B' \neq B$ be any other bounded subset with the property that, for every compact subset X_0 of X, there is a time $T > 0$ such that $\|x(t, x_0)\| \in B'$ for all $t \geq T$ and all $x_0 \in X_0$. Then, it is easy to check that $\omega(B') = \omega(B)$. Thus, in view of the properties described in Lemma 2 above, the following definition can be adopted (see [7]).

Definition Suppose the motions of system (6), with initial conditions in a closed and positively invariant set X, are ultimately bounded. A *steady state* motion is any motion with initial condition $x(0) \in \omega(B)$. The set $\omega(B)$ is the *steady state locus* of (6) and the *restriction* of (6) to $\omega(B)$ is the *steady state behavior* of (6).

The notion introduced in this way recaptures the classical notion of steady state for linear systems and provides a new powerful tool to deal with similar issues in the case of nonlinear systems.

Example In order to see how this notion includes the classical viewpoint, consider an n-dimensional, single-input, *asymptotically stable* linear system

$$\dot{z} = Fz + Gu \tag{7}$$

forced by the harmonic input $u(t) = u_0 \sin(\omega t + \phi_0)$. A simple method to determine the periodic motion of (7) consists in viewing the forcing input $u(t)$ as provided by an autonomous "signal generator" of the form

$$\dot{w} = Sw \qquad u = Qw$$

in which

$$S = \begin{pmatrix} 0 & \omega \\ -\omega & 0 \end{pmatrix} \qquad Q = \begin{pmatrix} 1 & 0 \end{pmatrix}$$

and in analyzing the state state behavior of the associated "augmented" system

$$\begin{aligned} \dot{w} &= Sw \\ \dot{z} &= Fz + GQw\,. \end{aligned} \tag{8}$$

As a matter of fact, let Π be the unique solution of the Sylvester equation $\Pi S = F\Pi + GQ$ and observe that the graph of the linear map

$$\begin{aligned} \pi : \mathbb{R}^2 &\to \mathbb{R}^n \\ w &\mapsto \Pi w \end{aligned}$$

is an invariant subspace for the system (8). Since all trajectories of (8) approach this subspace as $t \to \infty$, the limit behavior of (8) is determined by the restriction of its motion to this invariant subspace.

Revisiting this analysis from the viewpoint of the more general notion of steady state introduce above, let $W \subset \mathbb{R}^2$ be a set of the form

$$W = \{w \in \mathbb{R}^2 \colon \|w\| \leq c\} \tag{9}$$

in which c is a fixed number, and suppose the set of initial conditions for (8) is $W \times \mathbb{R}^n$. This is in fact the case when the problem of evaluating the periodic response of (7) to harmonic inputs whose amplitude does not exceed a fixed number c is addressed. The set W is compact and invariant for the upper subsystem of (8) and, as it is easy to check, the ω-limit set of W under the motion of the upper subsystem of (8) is the subset W itself.

The set $W \times \mathbb{R}^n$ is closed and positively invariant for the full system (8) and, moreover, since the lower subsystem of (8) is a linear asymptotically stable system driven by a bounded input, it is immediate to check that the motions of system (8), with initial conditions taken in $W \times \mathbb{R}^n$, are ultimately bounded. As a matter of fact, any bounded set B of the form

$$B = \{(w, z) \in \mathbb{R}^2 \times \mathbb{R}^n : w \in W, \|z - \Pi w\| \le d\}$$

in which d is any positive number, has the property indicated in the definition of ultimate boundedness. It is easy to check that

$$\omega(B) = \{(w, z) \in \mathbb{R}^2 \times \mathbb{R}^n : w \in W, z = \Pi w\} ,$$

i. e. $\omega(B)$ is the graph of the restriction of the map π to the set W. The restriction of (8) to the invariant set $\omega(B)$ characterizes the steady state behavior of (7) under the family of all harmonic inputs of fixed angular frequency ω, and amplitude not exceeding c.

Example A similar result, namely the fact that the *steady state locus* is the *graph* of a map, can be reached if the "signal generator" is any nonlinear system, with initial conditions chosen in a compact invariant set W. More precisely, consider an augmented system of the form

$$\begin{aligned} \dot{w} &= s(w) \\ \dot{z} &= Fz + Gq(w) , \end{aligned} \tag{10}$$

in which $w \in W \subset \mathbb{R}^r$, $x \in \mathbb{R}^n$, and assume that: (i) all eigenvalues of F have negative real part, (ii) the set W is a compact set, invariant for the the upper subsystem of (10).

As in the previous example, the ω-limit set of W under the motion of the upper subsystem of (10) is the subset W itself. Moreover, since the lower subsystem of (10) is a linear asymptotically stable system driven by the bounded input $u(t) = q(w(t, w_0))$, the motions of system (10), with initial conditions taken in $W \times \mathbb{R}^n$, are ultimately bounded.

It is easy to check that the steady state locus of (10) is the graph of the map

$$\begin{aligned} \pi : W &\to \mathbb{R}^n \\ w &\mapsto \pi(w) , \end{aligned}$$

defined by

$$\pi(w) = \lim_{T\to\infty} \int_{-T}^{0} \mathrm{e}^{-F\tau} Gq(w(\tau, w))\,\mathrm{d}\tau . \tag{11}$$

To see why this is the case, pick any initial condition (w_0, z_0) for (10) on the graph of π and compute the solution $z(t)$ of the lower equation of (10) by means of the classical variation of constants formula, to obtain

$$z(t) = \mathrm{e}^{Ft} z_0 + \int_0^t \mathrm{e}^{F(t-\tau)} Gq(w(\tau, w_0))\,\mathrm{d}\tau$$

Since by hypothesis $z_0 = \pi(w_0)$, using (10) one obtains

$$\begin{aligned} z(t) =& \mathrm{e}^{Ft} \int_{-\infty}^{0} \mathrm{e}^{-F\tau} Gq(w(\tau, w_0))\,\mathrm{d}\tau \\ &+ \int_0^t \mathrm{e}^{F(t-\tau)} Gq(w(\tau, w_0))\,\mathrm{d}\tau \\ =& \int_{-\infty}^{t} \mathrm{e}^{F(t-\tau)} Gq(w(\tau, w_0))\,\mathrm{d}\tau \\ =& \int_{-\infty}^{0} \mathrm{e}^{-F\theta} Gq(w(\theta + t, w_0))\,\mathrm{d}\theta \\ =& \int_{-\infty}^{0} \mathrm{e}^{-F\theta} Gq(w(\theta, w(t, w_0)))\,\mathrm{d}\theta \\ =& \pi(w(t, w_0)) = \pi(w(t)) \end{aligned}$$

which proves the invariance of the graph of π for (10). It is deduced from this that that any point of the graph of π is necessarily a point of the steady state locus of (10). To complete the proof of the claim it remains to show that no other point of $W \times \mathbb{R}^n$ can be a point of the steady state locus. But this is a straightforward consequence of the fact that F has eigenvalues with negative real part.

There are various ways in which the result discussed in the previous example can be generalized. For instance, it can be extended to describe the steady state response of a nonlinear system

$$\dot{z} = f(z, u) \tag{12}$$

in the neighborhood of a locally exponentially stable equilibrium point. To this end, suppose that $f(0, 0) = 0$ and that the matrix

$$F = \left[\frac{\partial f}{\partial z}\right](0, 0)$$

has all eigenvalues with negative real part. Then, it is well known (see e. g. p. 275 [13]) that it is always possible to find a compact subset $Z \subset \mathbb{R}^n$, which contains $z = 0$ in its interior and a number $\sigma > 0$ such that, if $\|z_0\| \in Z$ and $\|u(t)\| \le \sigma$ for all $t \ge 0$, the solution of (12) with initial condition $z(0) = z_0$ satisfies $\|z(t)\| \in Z$ for all $t \ge 0$. Suppose that the input u to (12) is produced, as before, by a signal generator of the form

$$\begin{aligned} \dot{w} &= s(w) \\ u &= q(w) \end{aligned} \tag{13}$$

with initial conditions chosen in a compact invariant set W and, moreover, suppose that, $\|q(w)\| \le \sigma$ for all $w \in W$. If this is the case, the set $W \times Z$ is positively invariant for

$$\begin{aligned} \dot{w} &= s(w) \\ \dot{z} &= f(z, q(w)) , \end{aligned} \tag{14}$$

and the motions of the latter are ultimately bounded, with $B = W \times Z$. The set $\omega(B)$ may have a complicated structure but it is possible to show, by means arguments similar to those which are used in the proof of the Center Manifold theorem, that if Z and B are small enough the set in question can still be expressed as the graph of a map $z = \pi(w)$. In particular, the graph in question is precisely the center manifold of (14) at $(0, 0)$ if $s(0) = 0$ and the matrix

$$S = \left[\frac{\partial s}{\partial w}\right](0)$$

has all eigenvalues on the imaginary axis.

A common feature of the examples discussed above is the fact that the steady state locus of a system of the form (14) can be expressed as the graph of a map $z = \pi(w)$. This means that, so long as this is the case, a system of this form has a *unique* well defined *steady state response* to the input $u(t) = q(w(t))$. As a matter of fact, the response in question is precisely $z(t) = \pi(w(t))$. Of course, this may not always be the case and *multiple* steady state responses to a given input may occur. In general, the following property holds.

Lemma 3 *Let W be a compact set, invariant under the flow of (13). Let Z be a closed set and suppose that the motions of (14) with initial conditions in $W \times Z$ are ultimately bounded. Then, the steady state locus of (14) is the graph of a set-valued map defined on the whole of W.*

Necessary Conditions for Output Regulation

Taking advantage of the notions introduced in the previous section, we are now in a position to highlight some general properties that any controller that solves a problem of output regulation must necessarily have. Recall that, as defined earlier, the problem of output regulation is solved if, in the composite system (5):

- the positive orbit of $W \times X \times \Xi$ is bounded,
- $\lim_{t\to\infty} e(t) = 0$, uniformly in the initial condition.

The notions introduced in the previous section are instrumental to prove the following, elementary – but fundamental – result, which is a nonlinear enhancement of a Lemma of [10] on which all the theory of output regulation for linear systems is based.

Lemma 4 *Suppose the positive orbit of $W \times X \times \Xi$ is bounded. Then*

$$\lim_{t\to\infty} e(t) = 0$$

if and only if

$$\omega(W \times X \times \Xi) \subset \{(w, x, \xi) \colon h(w, x) = 0\}. \qquad (15)$$

It is seen from this simple result that the problem of output regulation can be simply cast as the problem of *shaping the steady state locus of the closed loop system*, in such a way that property (15) holds.

To proceed with the analysis in a more concrete fashion, we consider from now on the special case in which the controlled plant (4) is modeled by equations *in normal form*

$$\begin{aligned}
\dot z &= f_0(w, z) + f_1(w, z, e_1)e_1\\
\dot e_1 &= e_2\\
&\vdots\\
\dot e_{r-1} &= e_r\\
\dot e_r &= q(w, z, e_1, \ldots, e_r) + b(w, z, e_1, \ldots, e_r)u\\
e &= e_1\\
y &= col(e_1, \ldots, e_r)\,,
\end{aligned} \qquad (16)$$

with state $(z, e_1, \ldots, e_r) \in \mathbb{R}^{n-r} \times \mathbb{R}^r$, control input $u \in \mathbb{R}$, regulated output $e \in \mathbb{R}$, measured output $y \in \mathbb{R}^r$. The functions $f_0(\cdot)$, $f_1(\cdot)$, $q(\cdot)$, $b(\cdot)$, $s(\cdot)$ in (16) and (3) are assumed to be at least continuously differentiable. It is also assumed that

$$b(w, z, e_1, \ldots, e_r) \neq 0 \qquad \forall (w, z, e_1, \ldots, e_r)\,.$$

The initial conditions of (16) range on a set $Z \times E$, in which Z is a fixed *compact* subset of $\mathbb{R}^{n-r}$ and $E = \{(e_1, \ldots, e_r) \in \mathbb{R}^r \colon |e_i| \leq c\}$, with c a fixed number.

Suppose that a controller of the form (4) solves the problem of output regulation. Then Lemma 4 applies and, since $e = e_1$, we deduce that the steady state locus of the closed loop system (5) is necessarily a subset of the set of all states in which $e_1 = 0$. This being the case, it is seen from the form of the equations (16) that, when the closed loop system (5) is in steady state, necessarily also

$$e_2 = e_3 = \cdots = e_r = 0\,.$$

As a consequence, the following conclusions hold:

- *The steady state locus $\omega(W \times Z \times E \times \Xi)$ of the closed-loop system is a subset of the set $\mathbb{R}^s \times \mathbb{R}^{n-r} \times \{0\} \times \mathbb{R}^\nu$.*
- *The restriction of the closed-loop system to its steady state locus $\omega(W \times Z \times E \times \Xi)$ reduces to*

$$\begin{aligned}
\dot w &= s(w)\\
\dot z &= f_0(w, z)\\
\dot \xi &= \varphi(\xi, 0)\,.
\end{aligned} \qquad (17)$$

- *For each* $(w, z, 0, \ldots, 0, \xi) \in \omega(W \times Z \times E \times \Xi)$

$$0 = q(w, z, 0, \ldots, 0) + b(w, z, 0, \ldots, 0)\gamma(\xi, 0)\,. \quad (18)$$

The prior analysis implicitly assumes that the positive orbit of W under the flow of exosystem is bounded, i. e. that the motions of the exosystem asymptotically approach the its own steady state locus $\omega(W)$. In principle, $\omega(W)$ may differ from W but there is no loss of generality in assuming from the very beginning that the two sets coincide. After all, the problem in question is a problem concerning how the closed-loop system behaves in steady state and there is no special interest in considering exosystems that are not "in steady state". We make this assumption precise as follows.

Assumption (i) *The compact set W is invariant for (3).*

With this in mind we observe that, by Lemma 3, if the positive orbit of $W \times Z \times E \times \Xi$ under the flow of (5) is bounded, then $\omega(W \times Z \times E \times \Xi)$ is the graph of a (possibly set-valued) map defined on the whole of W. Consider now the set

$$\mathcal{A}_{\mathrm{ss}} = \{(w, z)\colon (w, z, 0, \ldots, 0, \xi) \in \omega(W \times Z \times E \times \Xi), \\ \text{for some } \xi \in \mathbb{R}\}$$

and define the map

$$u_{\mathrm{ss}} : \mathcal{A}_{\mathrm{ss}} \to \mathbb{R}$$
$$(w, z) \mapsto -\frac{q(w, z, 0, \ldots, 0)}{b(w, z, 0, \ldots, 0)}\,.$$

By construction, the set $\mathcal{A}_{\mathrm{ss}}$ is the graph of a (possibly set-valued) map defined on the whole of W, which is invariant for the dynamics of

$$\begin{aligned} \dot{w} &= s(w) \\ \dot{z} &= f_0(w, z)\,, \end{aligned} \quad (19)$$

that are precisely *the zero dynamics of the "augmented system"* (3)–(16), while the map $u_{\mathrm{ss}}(\cdot)$ is the control that forces the motion of (3) –(16) to evolve on $\mathcal{A}_{\mathrm{ss}}$.

With this in mind, the conclusions reached above can be rephrased in the following terms. Suppose that a controller of the form (4) solves the problem of output regulation for (16) with exosystem (3). Then, there exists a (possibly set-valued) map defined on the whole of W whose graph $\mathcal{A}_{\mathrm{ss}}$ is invariant for the autonomous system (19). Moreover, for each $(w_0, z_0) \in \mathcal{A}_{\mathrm{ss}}$ there is a point $\xi_0 \in \mathbb{R}^\nu$ such that the integral curve of (19) issued from (w_0, z_0) and the integral curve of

$$\dot{\xi} = \varphi(\xi, 0)$$

issued from ξ_0 satisfy

$$u_{\mathrm{ss}}(w(t), z(t)) = \gamma(\xi(t))\,, \qquad \forall t \in \mathbb{R}\,.$$

This is a nonlinear version of the celebrated *internal model principle* of [11].

Sufficient Conditions for Output Regulation

The Control Structure

On the basis of the ideas presented in the previous section we proceed now with the construction of a controller that solves the problem of output regulation. The "steady state" features of this controller are those identified at the end of the section, namely this controller has to be able to "generate" all controls of the form $u_{\mathrm{ss}}(w(t), z(t))$ for any "steady state" trajectory $w(t), z(t)$ of (19). The controller should incorporate a device that generates all such trajectories (the *internal model*), thus making sure that the "appropriate" state-state behavior takes place, and a device guaranteeing that convergence to this specific steady state behavior occurs. It is here that additional assumptions are needed.

Note that, since W is invariant for $\dot{w} = s(w)$, the closed cylinder

$$C := W \times \mathbb{R}^{n-r}$$

is locally invariant for (19). Hence, it is natural regard (19) as a system defined on C and endow the latter with the subset topology.

Assumption (ii) *There exists a bounded subset B of C which contains the positive orbit of the set W × Z under the flow of (19) and the resulting omega-limit set*

$$\mathcal{A} := \omega(W \times Z)$$

satisfies

$$(w, z) \in C\,, \quad |(w, z)|_{\mathcal{A}} \le d_0 \quad \Rightarrow \quad z \in Z \qquad (20)$$

where d_0 is a positive number.

While in the analysis of the necessity we have only identified the existence of a compact set (actually, the graph of a map defined on W) which is invariant for (19), the new assumption (ii) implies, in its first part, the existence of a compact set $\mathcal{A}$ (still the graph of a map defined on W) which is not only invariant but also uniformly attractive of all trajectories of (19) issued from points of $W \times Z$. The second part of the assumption, in turn, guarantees that this set is also stable in the sense of Lyapunov. In the next assumption we strengthen this property by also requiring the set $\mathcal{A}$ to be locally exponentially stable (this assumption is useful to straighten the subsequent analysis, but is not essential).

Assumption (iii) *There exist $M \geq 1$, $\lambda > 0$ such that*

$$\begin{gathered}(w_0, z_0) \in C\,, \quad |(w_0, z_0)|_{\mathcal{A}} \leq d_0 \quad \Rightarrow \\ |(w(t), z(t))|_{\mathcal{A}} \leq M e^{-\lambda t} |(w_0, z_0)|_{\mathcal{A}}\,, \qquad \forall t \geq 0\end{gathered}$$

in which $(w(t), z(t))$ denotes the solution of (19) passing through (w_0, z_0) at time $t = 0$.

To simplify the exposition, we address the special case in which the controlled system (16) has relative degree 1, and in which the coefficient $b(w, z, e_1)$ is identically equal to 1. In other words, we consider a system modeled by equations of the form

$$\begin{aligned}\dot{z} &= f_0(w, z) + f_1(w, z, e)e \\ \dot{e} &= q(w, z, e) + u \\ y &= e\,.\end{aligned} \tag{21}$$

There is no loss of generality in considering a system having this simple form (21) because, as shown for instance in [9,22], the case of a more general system of the form (16) can easily be reduced, by appropriate manipulations, to this one.

For convenience, rewrite the augmented system (3)–(21) as

$$\begin{aligned}\dot{z} &= f_0(z) + f_1(z, e)e \\ \dot{e} &= q_0(z) + q_1(z, e)e + u\end{aligned} \tag{22}$$

having set $z = (w, z)$. Consistently let $Z := W \times Z$ denote the compact set where the initial condition $z(0)$ is supposed to range. In these notations, assumptions (i) – (ii) - (iii) express the property that, in the autonomous system

$$\dot{z} = f_0(z)\,, \tag{23}$$

the set $\mathcal{A}$ is asymptotically and locally exponentially stable, with a domain of attraction that contains the set Z.

Suppose now that the control u is chosen as $u = -ke$. The closed-loop system thus obtained can be regarded as a feedback interconnection of

$$\dot{z} = f_0(z) + f_1(z, e)e \tag{24}$$

viewed as a system with input e and state z, and

$$\dot{e} = q_0(z) + q_1(z, e)e - ke \tag{25}$$

viewed as a system with input z and state e.

By assumption, system (24) possesses, when $e = 0$, an invariant set $\mathcal{A}$ which is asymptotically and locally exponentially stable, with a domain of attraction that contains the set Z of all admissible initial conditions. Thus, standard arguments (see, e. g [6].) can be invoked to claim that, if k is large enough, all trajectories of the interconnection (24)–(25) with initial conditions in $Z \times E$ remain bounded and the state (z, e) can be steered to an arbitrary small neighborhood of the set $\mathcal{A} \times \{0\}$. This does not solve the problem at issue, though, because the variable $e(t)$ is not guaranteed to converge to zero (but only to converge to a neighborhood of zero, whose size can be made arbitrarily small by increasing the gain coefficient k). The condition for having $e(t) \to 0$ as $t \to \infty$ is simply that the "coupling" term $q_0(z)$ vanishes on the set $\mathcal{A}$, but there is no reason for this to occur (see again, e. g.[6]). This is why a more elaborate, internal-model-based, controller is needed.

System (16) being affine in the control input u, it seems natural to look for a controller having a similar structure, namely a controller of the form

$$\begin{aligned}\dot{\xi} &= \varphi(\xi) + Gv \\ u &= \gamma(\xi) + v\end{aligned} \tag{26}$$

with state $\xi \in \mathbb{R}^\nu$, in which v is a residual control input, to be eventually chosen as a function of the measured output y. Here $\varphi(\cdot)$, G and $\gamma(\cdot)$ are functions to be determined. We will show in what follows that, if the triplet $\{\varphi(\xi), G, \gamma(\xi)\}$ possesses what we will define as *asymptotic internal model* property, the choice of the residual control v in (26) as

$$v = -ke$$

solves the problem of output regulation, provided that the gain coefficient k is sufficiently high.

The Internal Model

Controlling this system by means of (26) yields a closed-loop system

$$\begin{aligned}\dot{z} &= f_0(z) + f_1(z, e)e \\ \dot{e} &= q_0(z) + q_1(z, e)e + \gamma(\xi) + v \\ \dot{\xi} &= \varphi(\xi) + Gv\end{aligned} \tag{27}$$

which, regarded as a system with input v and output e, has relative degree 1 and zero dynamics given by

$$\begin{aligned}\dot{z} &= f_0(z) \\ \dot{\xi} &= \varphi(\xi) - G[\gamma(\xi) + q_0(z)]\,.\end{aligned} \tag{28}$$

System (27) can be put in normal form by means of the change of variables

$$\chi = \xi - Ge$$

which yields

$$\begin{aligned}\dot{z} &= f_0(z) + f_1(z, e)e\\ \dot{\chi} &= \varphi(\chi + Ge) - G\gamma(\chi + Ge) - Gq_0(z) - Gq_1(z, e)e\\ \dot{e} &= q_0(z) + q_1(z, e)e + \gamma(\chi + Ge) + v\,.\end{aligned} \tag{29}$$

Setting $x = (z, \chi)$, this system can be further rewritten in the form

$$\begin{aligned}\dot{x} &= f(x) + \ell(x, e)e\\ \dot{e} &= q(x) + r(x, e)e + v\end{aligned} \tag{30}$$

in which

$$f(x) = \begin{pmatrix} f_0(z) \\ \varphi(x) - G[\gamma(x) + q_0(z)] \end{pmatrix}$$

$$q(x) = q_0(z) + \gamma(x)$$

and $\ell(x, e)$, $r(x, e)$ are suitable continuous functions.

Suppose now that the residual control v is chosen as $v = -ke$. This yields a closed-loop system having a structure similar to the one considered in the previous subsection, namely a feedback interconnection of

$$\dot{x} = f(x) + \ell(x, e)e \tag{31}$$

viewed as a system with input e and state x, and

$$\dot{e} = q(x) + r(x, e)e - ke \tag{32}$$

viewed as a system with input x and state e. As claimed earlier, a high-gain control on e, namely a control $v = -ke$ with large k, would succeed in steering $e(t)$ to zero if two conditions are fulfilled:

(P1) the dynamics (28) possesses a compact invariant set which is asymptotically (and locally exponentially) stable, with a domain of attraction that contains the set $Z \times \Xi$ of all admissible initial conditions, and

(P2) the function $q_0(z) + \gamma(\xi)$ vanishes on this invariant set.

These are the properties that will be sought in what follows. Note that the fulfillment of these is determined only by properties of the autonomous system (23) and of the function

$$\rho = q_0(z) \tag{33}$$

which, in the composite system (28), can be viewed as the output of (23) driving a system of the form

$$\dot{\xi} = \varphi(\xi) - G[\gamma(\xi) + \rho]\,. \tag{34}$$

For convenience, we will say that triplet $\{\varphi(\xi), G, \gamma(\xi)\}$ *is an asymptotic internal model of the pair* (23) – (33) if properties (P1) and (P2) are satisfied. In this terminology, we can summarize as follows the conclusion obtained so far.

Proposition 1 *Pick compact sets Z, E and Ξ for the initial conditions of the closed-loop system (3), (21), (26). Assume that (i)-(ii)-(iii) hold and that the triplet $\{\varphi(\xi), G, \gamma(\xi)\}$ is an asymptotic internal model of (23) – (33). Then there exists $k^\star > 0$ such that for all $k \geq k^\star$ the controller (26) with $v = -ke$ solves the generalized tracking problem.*

The notion of steady state provides a useful interpretation of the properties in question. In fact, recall that, by assumption, all trajectories of system (23) with initial conditions in Z asymptotically converge to the compact invariant set $\mathcal{A}$, and the latter is also locally exponentially stable. If property (P1) holds, all trajectories of the composite system (28) with initial conditions in $Z \times \Xi$ asymptotically converge to the limit set $\omega(Z \times \Xi)$. Since (28) is a triangular system, it is readily seen that the set $\omega(Z \times \Xi)$ is the graph of a set-valued map defined on $\mathcal{A}$, i. e. that there exists a map

$$\tau : z \in \mathcal{A} \mapsto \tau(z) \subset \mathbb{R}^\nu\,,$$

such that

$$\omega(Z \times \Xi) = \{(z, \xi) \colon z \in \mathcal{A}, \xi \in \tau(z)\} \colon = \mathrm{gr}(\tau)\,.$$

The set $\mathrm{gr}(\tau)$ is the steady state locus of (28) and the restriction of the latter to this invariant set characterizes its steady state behavior. Property (P2), on the other hand, expresses the property that at each point of $(z, \xi) \in \mathrm{gr}(\tau)$

$$q_0(z) = -\gamma(\xi)\,. \tag{35}$$

Thus, looking again at system (28), it is realized that $\mathrm{gr}(\tau)$ is in fact invariant for

$$\begin{aligned}\dot{z} &= f_0(z)\\ \dot{\xi} &= \varphi(\xi)\,.\end{aligned} \tag{36}$$

Note that, if the map $\tau(z)$ is single-valued and C^1, its invariance for (36) is expressed by the property that

$$\frac{\partial \tau(z)}{\partial z} f_0(z) = \varphi(\tau(z)) \qquad \forall\, z \in \mathcal{A}\,, \tag{37}$$

while the fact that (35) holds at each point of $(z, \xi) \in \mathrm{gr}(\tau)$ is expressed by the property that

$$q_0(z) = -\gamma(\tau(z)) \qquad \forall\, z \in \mathcal{A}\,. \tag{38}$$

The Design of an Internal Model

As we have seen in the earlier sections, the proposed controller, if the asymptotic internal model property holds, is able to force – in the closed loop system – convergence to a steady state in which the regulated variable is identically zero. As a consequence, the controller solves the generalized tracking problem. It remains to be shown, therefore, how the asymptotic internal model property can be obtained. To this end, it is convenient to observe that the properties required in (P1) and (P2) are quite similar to properties that are usually sought in the design of *state observers*. As a matter of fact it is seen from (37) and (38) that, for each $z_0 \in \mathcal{A}$, the function of time

$$\hat{\xi}(t) = \tau(z(t, z_0))$$

which is defined (and bounded) for all $t \in \mathbb{R}$ satisfies

$$\frac{\mathrm{d}\hat{\xi}(t)}{\mathrm{d}t} = \varphi(\hat{\xi}(t)) \tag{39}$$

and, moreover

$$\gamma(\hat{\xi}(t)) = -q_0(z(t, z_0)) \,.$$

In view of the latter, system (34) can be rewritten in the form

$$\dot{\xi} = \varphi(\xi) + G[\gamma(\hat{\xi}) - \gamma(\xi)] \tag{40}$$

and interpreted as *a copy of the dynamics* (39) *of* $\hat{\xi}$ *corrected by an "innovation term"* $[\gamma(\hat{\xi}) - \gamma(\xi)]$ *weighted by an "output injection gain"* G. This is the classical structure on an *observer* and the requirement in (P1) expresses the property that the difference $\xi(t) - \hat{\xi}(t)$ (the "observation error", in our interpretation) should asymptotically decay to zero (with ultimate exponential decay).

This interpretation is at the basis of a number of major recent advances in the design of regulators. In fact, in a number of recent papers, this interpretation has been pursued and, taking into consideration various approaches to the design of nonlinear observers, has lead to effective design methods (see [5,9,22]). Two of such design methods are highlighted in the remaining part of this section.

The High-Gain Observer as an Internal Model (see [5])

The construction summarized in this section relies upon the following additional hypothesis.

Assumption (iv) *Suppose there exist an integer $d > 0$ and a locally Lipschitz function $f\colon \mathbb{R}^d \to \mathbb{R}$ such that, for any $z_0 \in \mathcal{A}$, the solution $z(t)$ of passing through z_0 at time $t = 0$ is such that the function $\rho(t) := q_0(z(t))$ satisfies*

$$\rho^{(d)}(t) = f(\rho(t), \rho^{(1)}(t), \ldots, \rho^{(d-1)}(t))$$

for all $t \in \mathbb{R}$.

Let $\tau\colon W \times \mathbb{R}^{n-1} \to \mathbb{R}^d$ be the map defined as

$$\tau(z) := \mathrm{col}(q_0(z), L_{f_0} q_0(z), \ldots, L_{f_0}^{d-1} q_0(z)) \tag{41}$$

and let $f_c\colon \mathbb{R}^d \to \mathbb{R}$ be a C^1 function with compact support which agrees with $f(\cdot)$ on $\tau(\mathcal{A})$. Then, it easy to check that the properties indicated in (37) and (38) are fulfilled by choosing

$$\varphi(\xi) = \begin{pmatrix} \xi_2 \\ \vdots \\ \xi_d \\ f_c(\xi_1, \xi_2, \ldots, \xi_d) \end{pmatrix}, \qquad \gamma(\xi) = \xi_1 \,. \tag{42}$$

Comparing this construction with the earlier remarks we observe, in particular, that system

$$\begin{aligned} \dot{z} &= f_0(z) \\ \rho &= q_0(z) \end{aligned} \tag{43}$$

is *immersed into a system which is uniformly observable*, in the sense of [12] (even though system (43) might not have had such a property). It is precisely this that makes it possible to choose G in such a way that the property indicated in (P1) can be achieved.

As a matter of fact, the property in question is achieved by choosing

$$G = D_g \begin{pmatrix} c_0 \\ \vdots \\ c_{d-1} \end{pmatrix}$$

where $D_g = diag(g, g^2, \cdots, g^d)$, g is a design parameter, and the c_i's are such that the polynomial $\lambda^d + c_0\lambda^{d-1} + \cdots + c_{d-1} = 0$ is Hurwitz, as formally proved in Lemmas 1 and 2 of [5] to which the interested reader is referred for details.

It is worth noting that the assumption in question clearly covers the interesting (and widely addressed in the recent past literature, see [15]) case in which the function $f(\cdot)$ is linear, namely the case in which (43) is immersed into a linear observable system. In this case, although the choice indicated above is clearly still valid, a more direct way of designing the regulator is to use $f(\cdot)$ instead of $f_c(\cdot)$ in the definition of $\varphi(\xi)$, and simply choose G in such a way that $\dot{\xi} = \varphi(\xi) - G\gamma(\xi)$ is a stable linear system.

The Andrieu-Praly's Observer as an Internal Model (see [22]) In this subsection we exploit certain results of the theory presented in [1] to weaken (and, to some extent, suppress) the Assumption (iv) presented at the beginning of the earlier subsection.

Let $(F, G) \in \mathbb{R}^{d \times d} \times \mathbb{R}^{d \times 1}$ be a controllable pair and set

$$\varphi(\xi) = F\xi + G\gamma(\xi), \tag{44}$$

with $\gamma\colon \mathbb{R}^d \to \mathbb{R}$ a continuous function to be determined later. If this is the case, the composite system (28) becomes

$$\begin{aligned} \dot{z} &= f_0(z) \\ \dot{\xi} &= F\xi - Gq_0(z)\,, \end{aligned} \tag{45}$$

which is precisely a system of the form (9), considered in the Example. If the matrix F is Hurwitz, and we restrict z to belong to the set $\mathcal{A}$, this system has a well defined steady state behavior, which is the graph of the map

$$\tau(z) = \int_{-\infty}^{0} \mathrm{e}^{-Fs} G q_0(z(s,z))\mathrm{d}s\,. \tag{46}$$

As shown in Example, the graph in question is invariant for system (45), is asymptotically (and locally exponentially) stable and with a domain of attraction that coincides with $W \times \mathbb{R}^d$. Moreover, it can also be shown that there exists a number $\ell > 0$ such that, if the eigenvalues of F have real part which is less ℓ, the map (46) is C^1 (see, e. g. [22]). If this is the case, to say that the graph of (46) is invariant for (45) is equivalent to say that

$$\frac{\partial \tau}{\partial z} f_0(z) = F\tau(z) - Gq_0(z) \qquad \forall\, z \in \mathcal{A}\,. \tag{47}$$

This being the case, it is immediate to check that properties (P1) and (P2) will be fulfilled if a function $\gamma(\xi)$ can be found that renders (38) satisfied. As a matter of fact, bearing in mind (44), condition (37) becomes

$$\frac{\partial \tau}{\partial z} f_0(z) = F\tau(z) - G\gamma(\tau(z)) \qquad \forall\, z \in \mathcal{A}$$

which, if condition (38) holds, reduces to (47). This, shows that a triplet having the asymptotic internal model property can be found if a function $\gamma(\cdot)$ exists which satisfies (38). It is here that the dimension d of the pair (F, G) plays a role, as formalized in the next proposition whose proof can be found in [22].

Proposition 2 *Suppose*

$$d \geq 2(s + n - r) + 2\,.$$

Then for almost all choices (see [22] for details) of a controllable pair (F, G), with F a Hurwitz matrix whose eigenvalues have real part which is less than ℓ, the map (46) satisfies

$$\tau(z_1) = \tau(z_2) \quad \Rightarrow \quad q_0(z_1) = q_0(z_2)\,.$$

As a consequence there exist a continuous map $\gamma\colon \tau(\mathcal{A}) \to \mathbb{R}$ such that

$$q_0(z) = -\gamma(\tau(z)) \qquad \forall\, z \in \mathcal{A}\,. \tag{48}$$

The map $\tau(\cdot)$ in (46) is defined only on $\mathcal{A}$, but is not difficult to extend it to a C^1 map defined on the whole set $W \times \mathbb{R}^{n-r}$, as shown in [22]. Also the map $\gamma(\cdot)$ that makes (48) true can be extended to the whole $\mathbb{R}^d$, but this extension is only known to be continuous.

We have shown in this way that the existence of triplet $\psi(\xi), G, \gamma(\xi)$ which has the internal model property can always be achieved, so long as the integer d is large enough. This result shows that the general design procedure outlined earlier in the article is always applicable (so long as the standing hypotheses (i)–(ii)–(iii) are applicable). From the constructive viewpoint, though, it must be observed that the result indicated in Proposition PR1 is only an existence result and that the function $\gamma(\xi)$, whose existence is guaranteed, is only known to be continuous. Obtaining continuous differentiability of such $\gamma(\xi)$ and a constructive procedure are likely to require further hypotheses, which should be in any case weaker than Assumption (iv) considered earlier, whose study is subject of current investigation.

Future Directions

One of the basic hypotheses of the theory described in the previous sections is that the zero dynamics of the augmented system consisting of the controlled plant and of the exosystem possess a compact attractor which is also locally asymptotically stable. This assumption, with an acceptable abuse of terminology, is usually referred to as the "minimum-phase" property. Another standing hypothesis is that the regulated variable coincides with the measured variable. The future directions of the research in this area are aimed at the removal of these assumptions. There are several directions in which the problem can be tackled. One way is the use of control structures by means of which the controlled system (plus a part of the controller) can be interpreted with systems having a different zero dynamics (in particular a zero dynamics possessing a compact attractor). This is the nonlinear equivalent of certain design procedures for linear systems based on the assignment of zeros. This technique has proven to be powerful in the stabilization of certain classes of nonlinear systems (see [18]) and is expected to be successful, with appropriate enhancements, in the design of regulators. The analysis of the necessary conditions for output regulation has also shown that, if the generalized tracking problem is solvable, the inverse dynamics of the augmented system consisting of the controlled plant and of the exosystem possesses a compact invariant set to which all initial conditions are asymptotically controllable (by means of the regulated variable viewed as a control). This set is not necessarily asymptotically stable (as it would be under the "minimum-phase" hypothesis) but could be asymptotically stabilizable (by either full-state or measurement-based feedback). Thus, another direction in which the research will evolve, in the development of control schemes for possibly non "minimum-phase" nonlinear systems, is based on the exploitation of the (weaker) assumption that in the zero dynamics of the augmented system there is a compact set that can be made invariant and asymptotically stable by means of feedback. This will yield a design procedure in which a virtual control (either full-state or measurements-based) is designed to stabilize that compact invariant set. The (possibly dynamic) controller obtained in this way will then be embedded into a regulator designed according to the principles developed in this article.

Bibliography

1. Andrieu V, Praly L (2006) On the existence of a Kazantis-Kravaris/Luenberger observer. SIAM J Contr Optim 45:432–456
2. Birkhoff GD (1927) Dynamical systems. American Mathematical Society, Providence
3. Byrnes CI, Delli Priscoli F, Isidori A, Kang W (1997) Structurally stable output regulation of nonlinear systems. Automatica 33:369–385
4. Byrnes CI, Isidori A (2003) Limit Sets, Zero dynamics and internal models in the problem of nonlinear output regulation. IEEE Trans Autom Contr 48:1712–1723
5. Byrnes CI, Isidori A (2004) Nonlinear internal models for output regulation. IEEE Trans Autom Contr 49:2244–2247
6. Byrnes CI, Isidori A, Praly L (2003) On the asymptotic properties of a system Arising in non-equilibrium theory of output regulation, preprint of the Mittag-Leffler Institute. 18, Stockholm
7. Isidori A, Byrnes CI (2007) The steady-state response of a nonlinear system: ideas, tools and applications, preprint
8. Davison EJ (1976) The robust control of a servomechanism problem for linear time-invariant multivariable systems. IEEE Trans Autom Contr AC-21:25–34
9. Delli Priscoli F, Marconi L, Isidori A (2006) A new approach to adaptive nonlinear regulation. SIAM J Contr Optim 45:829–855
10. Francis BA (1977) The linear multivariable regulator problem. SIAM J Contr Optim 14:486–505
11. Francis BA, Wonham WM (1976) The internal model principle of control theory. Automatica 12:457–465
12. Gauthier JP, Kupka I (2001) Deterministic observation theory and applications. Cambridge University Press, Cambridge
13. Hahn W (1967) Stability of motions. Springer, New York
14. Hale JK, Magalhães LT, Oliva WM (2002) Dynamics in infinite dimensions. Springer, New York
15. Huang J, Lin CF (1994) On a robust nonlinear multivariable servomechanism problem. IEEE Trans Autom Contr 39:1510–1513
16. Huang J, Rugh WJ (1990) On a nonlinear multivariable servomechanism problem. Automatica 26:963–972
17. Isidori A (1995) Nonlinear Control Systems. Springer, London
18. Isidori A (2000), A tool for semiglobal stabilization of uncertain non-minimum-phase nonlinear systems via output feedback. IEEE Trans Autom Contr AC-45:1817–1827
19. Isidori A, Byrnes CI (1990) Output regulation of nonlinear systems. IEEE Trans Autom Contr 25:131–140
20. Isidori A, Marconi L, Serrani A (2003) Robust autonomous guidance: An internal model-based approach. Springer, London
21. Khalil H (1994) Robust servomechanism output feedback controllers for feedback linearizable systems. Automatica 30:587–1599
22. Marconi L, Praly L, Isidori A (2006) Output stabilization via nonlinear luenberger observers. SIAM J Contr Optim 45:2277–2298
23. Serrani A, Isidori A, Marconi L (2001) Semiglobal nonlinear output regulation with adaptive internal model. IEEE Trans Autom Contr 46:1178–1194
24. Teel AR, Praly L (1995) Tools for semiglobal stabilization by partial state and output feedback. SIAM J Control Optim 33:1443–1485
25. Willems JC (1991) Paradigms and puzzles in the theory of dynamical systems, IEEE Transaction on Automatic Control 36:259–294

Systems Biology of Human Immunity and Disease

JARED C. ROACH
Seattle Childrens Hospital, Seattle, USA

Article Outline

Glossary

Complex disease A disease that has an etiology inconsistent with the simple models of genetic inheritance proposed by Gregor Mendel. A synonym for "complex disease" is "non-Mendelian disease." Type 2 diabetes is an example of a complex disease, as are most common diseases. The term "complex phenotype" is employed analogously to describe non-disease properties of individuals, such as height. Typically, complex diseases are influenced by multiple genetic and environmental factors. For most diseases, little is known about either of these sets of factors.

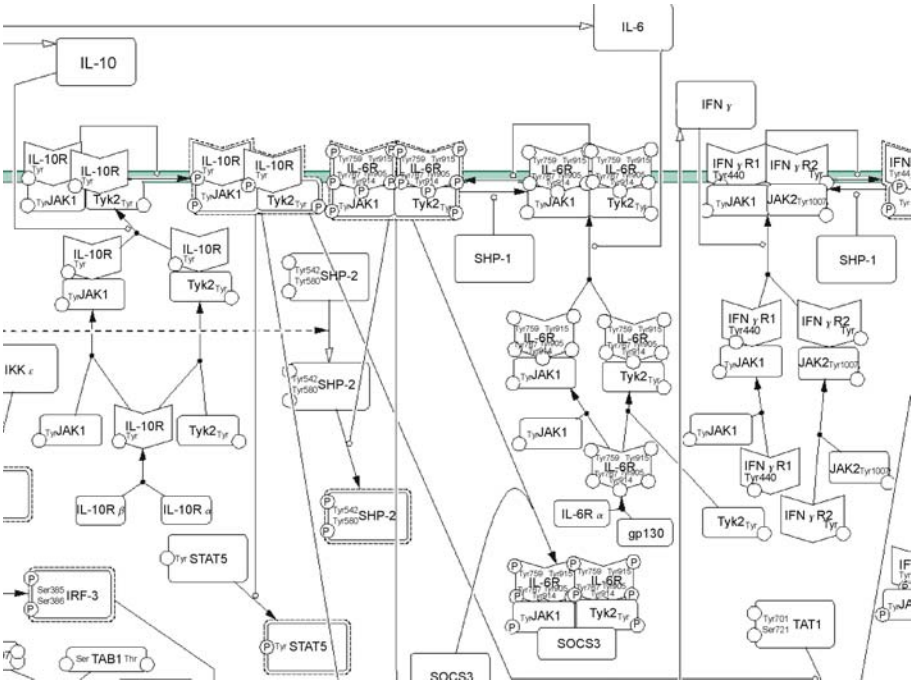
Nuclear regulation Many signals are integrated and processed on segments of DNA that are near or adjacent to genes. These segments of DNA are called cis-regulatory elements. Proteins that bind cis-regulatory elements are called transcription factors. The informational output of cis-regulatory signal integration is the rate of production of messenger RNA (mRNA) encoded by that gene. A major goal for many human systems biology projects is to understand key nuclear regulatory networks.

Regulatory network A paradigm for modeling information flow within a biological system, such as a cell. Elements of the system that can have multiple states are encoded as nodes; modes of communication between these elements are encoded as edges. A typical node would be a protein that might have multiple states related to its level of phosphorylation or cellular location. A typical edge would represent the catalytic effect of one protein upon another's state of phosphorylation.

Signaling pathway Signals typically reach cells through the binding of molecules to receptors on the cell surface. Over the course of seconds to minutes, the receptor interacts with other proteins or small molecules, changing their structure, concentration, or intracellular location. These changes can in turn cause other changes. One common ultimate effect is an alteration in transcription factor binding to target genes, in turn causing a change in mRNA levels. The term "signaling pathway" pre-dates modern systems biology, and represents a paradigm for information flow that is largely unidirectional and linear. Systems biology replaces this paradigm with that of a "regulatory network."

Definition of the Subject

Systems biology is the derivation of emergent properties of a multicomponent biological system through the construction of quantitative predictive models. These predictive models are typically formulated as networks. A portion of such a model is shown in Fig. 1. Human systems biology is the practice of systems biology to elucidate emergent properties of humans or any subcomponent of the human body, such as a cell. Human systems present challenges and opportunities not found in other organisms.



Systems Biology of Human Immunity and Disease, Figure 1

Example of a portion of a network model of a macrophage signaling network. This network was drafted with aid of CellDesigner software [34], and is reproduced with permission of The Systems Biology Institute (period)

Most problems in human systems biology are motivated by a desire to understand, predict, prevent, ameliorate, or cure a human disease. The human immune system is a mediator for most human diseases, and is the key system mediating autoimmune and infectious diseases. Many problems in human systems biology focus on the immune system.

As defined, human systems biology is a subset of systems biology. The subject merits special consideration largely because of medical importance. The human subset of systems biology is not an exceptionally distinct subset compared to other reasonable subsets, such as mammalian systems biology. However, it is fairly distinct from systems biology as applied to unicellular and invertebrate organisms.

Introduction

The difference between human systems biology and the practice of systems biology in other organisms is one of philosophy and emphasis [1]. The need to improve human health drives human systems biology. The emphasis of general systems biology is the development of technological, methodological, and algorithmic approaches to science. Human systems biology exploits these developments, but tends to be more focused on clinical endpoints. It is acceptable to approach these endpoints with incomplete, partial, hybrid, or modified versions of systems biology. Because these approaches include elements of systems biology, such as the analysis of high-throughput data, they are labeled as systems biology, even though they may not include all elements standard to general systems biology, such as the use of multiple orthologous data sets [5].

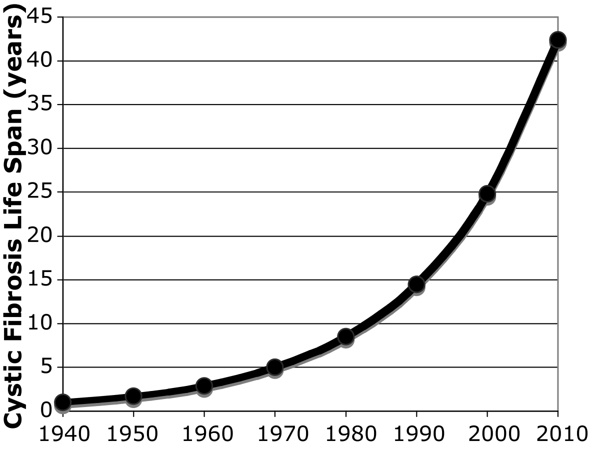
The needs of modern medicine set the context for human systems biology. From a data-driven point of view, there have been three major epochs of medicine: (1) observational medicine, (2) evidence-based medicine, and (3) predictive and personalized medicine [51]. These categories overlap to some extent, but they correspond fairly well to distinct periods of time. Observational medicine was prevalent from the dawn of civilization until early in the twentieth century. It is characterized by personal experience and oral tradition. An example and highlight of observational medicine was the discovery of digitalis as a treatment for dropsy, now known as edema. The discovery is credited in 1776 to William Withering, an English doctor who lived outside of London. Dr. Withering noticed that a patient with dropsy had been cured by a folk practitioner using a preparation from the foxglove plant. It is likely that this discovery had been made many times prior to the mid-eighteenth century, but lack of organized methods for communication and data handling kept the knowledge fragmented and incomplete. Dr. Withering was rich and connected. He was a member of the Lunar Society, a group of scientists founded by Matthew Boulton and Erasmus Darwin, the grandfather of Charles. These connections, foreshadowing those available to modern scientists through the internet, enabled the knowledge of foxglove to be preserved, distributed, and improved.

By the early twentieth century, incremental improvements to medical knowledge became increasingly common, and evidence-based medicine flourished. Advancements were enabled primarily by: (1) advances in communication, such as inexpensive publishing by printing presses, (2) chemistry, permitting the purification of active ingredients from botanical medicines, and (3) the nascent field of statistics. Statistics enabled legitimacy and falsifiability of medical facts. The first major application of statistics to medicine was the work of John Snow in 1843 to correlate water sources with cholera outbreaks. This correlation, and many others since, enabled preventive medicine. Chemistry, and later other fields of science and engineering, enabled reductionism. For example, reductionism enabled the identification of digitalis as the component of foxglove responsible for its effects on the cardiovascular system. The culmination of evidence-based medicine was the Framingham Heart Study circa 1948. Thousands of people from the town of Framingham and their descendants were observed longitudinally. Many health parameters were recorded. The Framingham study provided much of our modern knowledge concerning heart disease risk, including effects of diet, exercise, and aspirin. The Framingham Heart Study gave birth to the term "risk factor" and bolstered the recognition of prevention as the most cost-effective form of medicine. Prior to preventive medicine as a philosophy, a heart attack would be viewed as the beginning of a relationship between patient and doctor. With preventive medicine, the relationship would begin at birth or even before, and focus on modifying risk factors to prevent an attack from ever occurring. Data processing and analysis for such studies became increasingly sophisticated and required computers. Progress in evidence-based medicine was linked to the simultaneous revolution in computational power. In particular, the Framingham Heart Study benefited from scientific and sociological momentum generated in Word War II. Operations Research was largely born in the Battle of the Atlantic. The US population was willing to participate in large studies with society as the primary beneficiary. Command-and-control bureaucracies existed to coordinate large projects. Projects requiring the collaboration of hundreds of medical professionals were favorably viewed by researchers and their funding agencies.

By the end of the Human Genome Project at the beginning of the twenty-first century, limitations of the previous century's approach to evidence-based medicine were beginning to be recognized and solutions to these limitations were envisioned. Statistics, exemplified by linear regression models and chi-squared tests, required increasingly large cohorts to identify weaker risk factors. Knowledge gained from large cohorts was applicable only to populations (such as "all white males") and not individuals (such as the person reading this article). Studies became very expensive and could take decades or even centuries to elucidate correlations between causes and effects. The stage was set for the era of predictive and personalized medicine. A major approach to developing knowledge for this era of data-driven medicine has been and will be human systems biology.

In 1892, William Osler observed, "If it were not for the great variability among individuals medicine might be a science, not an art." Historically, the art of medicine was to personalize by trial and error. In the absence of personalization, a patient was prescribed a treatment based on population-based statistics. This is still the norm. If the treatment fails, the next best treatment from a population-inferred list will be prescribed. This process will repeat until the patient or doctor tires of it or dies, money becomes limiting, or effective treatment is achieved. Personalized medicine aims to match the best treatment for an individual without such perilous trial-and-error—and to turn Dr. Osler's art into a science. Previously, data for personalized models was based on orally conveyed family history, together gender and a crude estimate of genetic heritage, or race. Such personalization was tailored to subpopulations of people. 21st-century personalized medicine will develop predictions from an individual's genes and personal developmental and exposure history. Interventions will be tailored to that individual, not to a subpopulation to which the individual belongs. Systems-biology research creates the data and knowledge required by this approach. In particular, models are created based on the understanding of molecular mechanisms. Emergent properties arise from holistic analysis of the interactions in the modeled system. Each individual can be modeled based on their genetic risks, environmental exposures and their responses, biomarkers, and health assessments. Recommendations for particular medications, doses, interventions, and lifestyle changes will be personalized [12].

Expectations for modern research may be portrayed very highly in the popular press, with hopes for miracle cures. These may arise, but the most profound changes in medicine will most likely come from steady incremental advances. Research in all epochs of medicine has produced incremental advances. Cumulatively, these advances have enabled vast gains in human health (Fig. 2). Future incremental advances are expected throughout the epoch of personalized and preventive medicine begins.



Systems Biology of Human Immunity and Disease, Figure 2
Exponential rise in expected life span for cystic fibrosis. A fairly low exponent, 5.5% in this case, can lead to dramatic advances in human health over time. The goals of human systems biology are to drive such exponential advances

The phrase "systems biology" was first used early in the twentieth century to describe analyses of ecosystems, but the phrase never became popular until the turn of 21st century with the establishment of several institutes and research groups focused on systems-biology research and using the term "systems biology" to describe that research. The timing of this surge in interest was almost entirely due to the completion of the Human Genome Project and the need to develop research methodologies appropriate for the post-genomic era. The complete genome sequence provided a near-comprehensive "parts list" of genes. Coinciding approximately with the Human Genome Project, high-throughput technologies for measuring other important parts of human systems were developed, including technologies for measuring proteins and transcripts [18]. Neologisms with the suffix "omics" describing these new fields of analysis have become popular, including "proteomics" and "transcriptomics." Other factors driving modern systems biology have been (1) the tremendous boost to human collaboration and information sharing provided by the internet and (2) the conceptualization of biology as an information science. The digital nature of the genome has led to a belief that all of the individual components and interactions of biological systems are ultimately knowable, can be represented in data structures, and are amenable to computational analysis. The elucidation of the digital genome is a triumph of reductionism that has energized the pursuit of emergent properties. In the near future, complete sequences of the genomes of many humans will be known, further facilitating correlations between genotype and phenotype.

Modern systems biology focuses on systems at the cellular or subcellular level because systems of these complexities are considered to be neither too trivial to study nor too complex to comprehend. Some systems biology projects involve systems that are either more or less complex. In particular, human systems biology tends to focus on larger, more complex systems. For most problems of medical interest, it is generally anticipated that clinically useful predictions will require modeling at the multi-cellular level or at the level of the entire organism. The need to focus on very complex systems and to make clinically useful predictions underlies fundamental differences between the conduct of systems biology in humans and in other organisms.

There are tens of thousands of genes, small molecules, and proteins in a typical cellular system. Properties of these components and of the entire system need to be studied under many different conditions and dynamically over time periods ranging from seconds to years. There are several important generalizations that can be made about the techniques required to study such systems. Systems biology requires: (1) the analysis of high-throughput data sets, (2) the analysis of multiple qualitatively distinct different types of data ("orthologous data"), (3) a multidisciplinary collaborative team capable of both data generation ("bench work") and data analysis ("computational biology"). Also, because systems biology makes predictions, its practice is necessarily iterative: a model is built, predictions are tested, discrepancies between predictions and measurements are used to refine the model, predictions of the new model are tested, and so on until the incremental utility of the model refinements becomes small [47].

The application of the above approach to science is not new; it is merely new to biology. Other disciplines have used collaborative research employing multiple high-throughput data sources to produce models and emergent predictions. For example, in meteorology, a daily weather report is a simple prediction derived from such a process. One of the utilities of the phrase "systems biology" is that it brings attention to the methodological differences between this approach and the more prevalent reductionist approaches in biology during the twentieth century. The coining of the phrase thus helps catalyze a sociological change in biology.

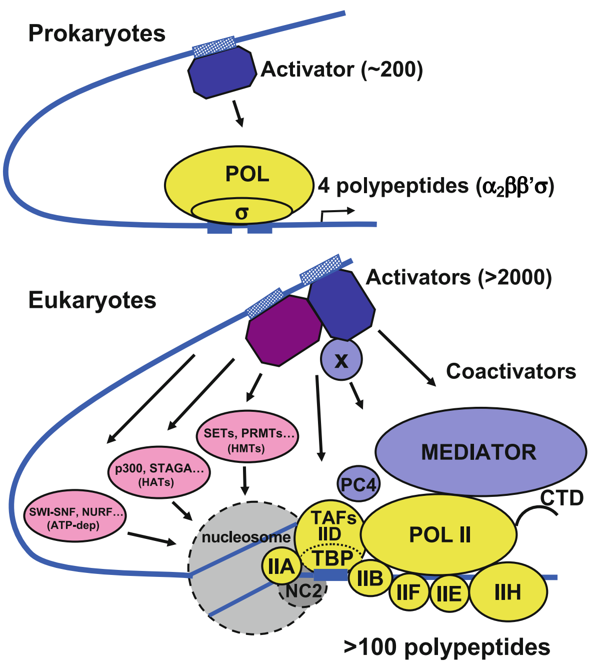
A number of disciplines that might semantically be considered to fit the definition of systems biology are not usually implied by users of the phrase because these disciplines were well established prior to the completion of the Human Genome Project. These include ecology, physiology, and biochemistry. In these fields, research may involve formulation of hypotheses as networks, and prediction of emergent properties of multicomponent systems [26].

Challenges and Solutions for Human Systems Biology

Human systems biology requires asking questions that are distinct from typical questions asked of prokaryotic systems. These distinct questions require distinct methodologies to answer [30].

There are many challenges for human systems biology. Underlying network models are far more complex than those of prokaryotes [43]. There are more components, or network nodes, of nearly every type. For example, there are more genes, as discussed in the accompanying chapter on genome complexity. Each gene has more cis-regulatory elements. On average, any given gene has more functions and products. These products are primarily mRNAs. On average, any given mRNA has more functions and products. These products are primarily proteins. Eukaryotic cells have more cellular compartments, including mitochondria, lysomomes, phagosomes, a nucleus, endoplasmic reticulum, and secretory granules. Humans have many more cells, and more cell types, and they are organized into different tissues [6]. Furthermore, there is no guarantee that insights gained from one human tissue will be easily extrapolatable to other human tissues; the same protein may play a very different role. In addition to many more nodes in human networks, there are many more interactions. In particular, the number and complexity of interactions operating on cis-regulatory elements of eukaryotic genes is greater (Fig. 3) [42,48].

Not only are human systems much more complex than other systems, but it is much harder to acquire data from these systems. Many experiments possible in other organisms are impossible in humans due to ethical concerns. Such concerns prevent the intentional creation of humans with identical genomes. Humans have very long life spans that essentially rule out acquiring data over the life span of particular individuals, particularly for systems biology, which requires iteration. Projects that might take weeks in a prokaryotic context would take many centuries if the experimental strategies were directly applied to human problems. The cost of performing experiments on human



Systems Biology of Human Immunity and Disease, Figure 3
Eukaryotic versus prokaryotic cis-regulation. The number of factors involved in regulating expression of a eukaryotic gene is at least an order of magnitude greater than for a prokaryotic gene. The complexity of the interactions is also greater. Reprinted from [42]

systems also tends to be much higher than in other systems.

As a consequence of the difficulty of research in mammals, whether that research is reductionist or holistic, on average for any given node or edge of a human network there is less prior information available than there would be for a simple model organism. In many cases, a literature or database search for information pertaining to a human gene will yield either no information or information restricted to a few cell types under only a few conditions [11]. For example, the target genes and sequences of most human transcription factors are completely unknown, leading to substantial uncertainties in models of human nuclear regulatory networks.

There are many solutions to these challenges. Some solutions remain visionary or require technologies that are in development, such as the ability to sequence an individuals genome for less than $1,000. If the sole goal of human research was knowledge for the sake of knowledge, then the most economically efficient approach to most human systems biology problems would be to wait for these technologies to be fully developed before applying them directly to disease research. However, medicine carries a personal urgency not found in most other areas of research. Delays in research can be measured in human life and well being. So in addition to investing in technology and methodology, human systems biology must work with the tools at hand.

Because so much less is known about the basic elements of human systems, a continued focus on the individual nodes and edges is important. Human systems biology projects do not displace traditional reductionist biology projects, but rather rely on them. A key fundamental question to be asked of all possible nodes in a network is that of membership: "Does this node belong in the network?" The same question can be asked of each possible edge. The various qualities of the network components also need to be determined. These include directionality of directed edges and rate constants for parts of the network modeled dynamically. If the most important nodes and edges in a network can be determined, future reductionist research can be appropriately directed to expand information on these network components. The complexity of human systems, the difficulty of automatically parsing the literature, and the high false positive rate in many human high-throughput data sets demands increased expert curation at all steps in the process. The sheer volume of the data makes such a call for human curation seemingly futile, but currently without such curation, most models will not make useful medical predictions. Appropriately directed projects will employ automated tools to reduce the burden of curation. For this purpose some of the most useful tools will be interfaces that present curators with information in forms that are easy to comprehend and integrate manually and that also permit decisions and rationales to be quickly recorded. The visual computer interfaces imagined for the 2002 science fiction movie "Minority Report" serve to stimulate ideas of what such interfaces might become.

Human systems biologists must also model simpler, or smaller, networks than would be appropriate for a simple model organism. Although somewhat paradoxical, this need arises from the intense need for data curation and the difficulty of acquiring high-quality data. The more complex the systems, the greater the number of free parameters. For larger networks, the lack of experimental data prevents these parameters from being constrained, and so no statistical confidence can be placed in predictions. Therefore, a reasonable approach for building a predictive network in a human system is to start very small. In some cases, networks with as few as two nodes and a single connecting edge can form a reasonable nucleus for the iterative process of building a systems biology network. Although such a network would be too trivial to discuss in a prokaryotic setting, it might be capable of making a clinically useful prediction and so would merit consideration in a human disease setting.

One possible approach to reducing the network complexity is to focus on only a few key nodes. Another is to model systems at grosser levels [9]. Thus rather than representing each protein as a node, one might group a set of proteins as a conceptual "module," and only represent the module as a node in the network. However, this approach requires exceptional care and curation, as it becomes hard to define and even harder to measure quantitative properties of modules [30]. One approach to developing hypotheses related to the functionality of modules is to perform comparative network analysis between different vertebrates [8]. Such analysis requires at least a moderate understanding of a network in at least two systems, which is currently rare, but may become more available over the next few decades.

It is easy for researchers to be distracted by the vast quantity of high-throughput data available, and to yield to reporting analyzes that are the result of automated in silico predictions of algorithms that often cannot be tested against gold standards. Human curation is often viewed as subjective, as it is hard to provide numerical statistics associated with such curation. Numerical statistics, on the other hand, are fairly easy to provide with output from automated algorithms, and these outputs are considered objective. From a medical point of view, however, the criterion of utility must trump evaluation of research based on simple ratio of subjectivity to objectivity. Medical objectivity must be evaluated at a deeper level of objectivity: quantitative measures of improved health care. Frequently, the output of automated algorithms is indeed objective, but is also nonsense. Examples of such nonsense are often manifested as very large networks consisting of hundreds or thousands of nodes and edges. Such networks are said to resemble hairballs and have been termed "ridiculomes." These networks have little if any predictive capability [32]. A weather report, as published in a daily newspaper, offers an example of a prediction in a useful form. Systems biology predictions should aim to be more like weather reports and less like hairballs.

Despite the large number of obstacles facing human research, there are a few paths to knowledge that work very well in humans. These derive largely from the ability of human research subjects to intellectually contribute to a research project. This is especially true for phenotypes of higher cognitive function. The most successful examples of human-specific research projects are family and genome-wide association studies. Volunteers with specific phenotypes bring themselves to the attention of researchers, permitting the genetic difference between affected and unaffected individuals to be determined. This is the most powerful and productive method that human systems biologists currently have for identifying the genes that are key nodes in human disease systems.

Given the fairly large number of differences in the state of knowledge and in the constraints on how research might be performed, the genre of questions asked by human systems biologists differs markedly from questions asked in other systems. Whereas design of a synthetic microorganism from the ground up might be a reasonable experimental agenda for a microbial biofuel project, other questions are more natural in human medicine.

Human-Specific Problems

It became fairly clear by the height of the era of evidence-based medicine that prevention has the highest ratio of benefit versus cost to society. Prevention is most effective when coupled to prediction. Interventions have risks and costs, so should be targeted to those predicted to have sufficient risk to benefit from a preventive intervention. Therefore a major goal of systems biology is to develop personalized predictive risk assessments. These risk assessments can be absolute (e. g., an individual's risk at the time of birth or at the time of a clinic visit) or conditional (e. g., risk if the individual stopped smoking). The types of outcomes that are useful to predict for an individual include whether or not a vaccination is protective, and whether a disease might occur and when it is likely to do so. A particular world health challenge today is the unpredictability of the effectiveness of the BCG vaccine for tuberculosis. Individuals who receive the vaccine have little way of knowing whether or not they will develop protective immunity [3]. The identification of biomarkers enabling such a prediction would be useful for counseling individuals as well as developing society-wide measures for preventing tuberculosis. Predictions at the level of a cell or tissue include what developmental programs these cells might follow or specific outcomes such as apoptosis, endocytosis, or mobility.

Other types of questions that can be asked in human systems include:

1. The identification of key information-transducing nodes in a system, as well as the inference of nodes that must be part of the system but are currently unsuspected or unknown. Such inference would be based on experimental results that differed from predicted results.
2. The identification of therapeutic drug targets [19].

3. Prediction of the effects of a gene knockout or knockdown, or overexpression of that gene.
4. Most generally, prediction of the effects of any perturbation to a system, including chemical and environmental perturbations.
5. Prediction of a particular property of a node or edge, such as Gene Ontology category, the presence of negative regulation, or the presence of positive or negative feedback.
6. Prediction of the concentration or flux or production of a specific gene, mRNA, or protein under a specific set of conditions at a specific time in a system.

The most successful projects will define exactly what questions will be addressed by the project before the project begins, and will define measurable quantitative variables that will be predicted by models [23]. The ability of a research group to precisely and accurately make quantitative predictions is one useful measure of whether a particular research methodology and approach is working effectively. Some questions lend themselves more naturally to quantitative predictions, such as those related to flux or concentration of small molecules. Others, such as the identification of drug targets, are very difficult to evaluate on any timescale less than a few decades.

Systems biology must be applied to all human diseases, as the urgency of clinical discovery is paramount. However, from the point of view of leveraging human systems biology research to gain fundamental understanding of a system, some human diseases may lend themselves more readily to knowledge discovery. These will be the diseases for which the etiology lies contained in a fairly small network and can be understood by understanding the interactions of a relatively few nodes. Thus, one might predict that research with Noonan Syndrome, which appears to be related to defects within the Ras signaling network may lead more immediately to fundamental understanding than research in type 1 diabetes, which appears to have etiologic effects spread throughout many cell types that arise from both genetic and environmental influences, and that act at disparate timepoints throughout the life of the individual.

Examples

Research in Noonan Syndrome represents an elegant example of systems biology as applied to human disease. Phenotypes of Noonan Syndrome include congenital heart malformation, short stature, and learning problems. Approximately 1 in 2,000 children are born with Noonan Syndrome. The range and severity of features can vary greatly between individuals, suggesting that there are multiple causes of the syndrome that affect common etiologic pathways, but in subtly different manners. Frequent transmission from parent to child suggested a genetic defect with autosomal dominant inheritance. Traditional genetic approaches mapped Noonan Syndrome to chromosome 12q24. By 2001, it was shown that about half of all individuals with Noonan syndrome were caused by a mutation of the PTPN11 gene at position 12q24. The PTPN11 gene encodes a tyrosine phosphatase. This discovery provided the first insight into the formation of a predictive model for Noonan Syndrome. With an underlying hypothesis that Noonan Syndrome was the emergent result of the output of an unknown dysfunctional network, the discovery that PTPN11 was part of that network allowed orthologous data sets and analysis to be leveraged [28,29]. PTPN11 is a node in an intracellular signaling network, including the traditional Ras/MAP kinase signaling pathway. This network governs cell division and differentiation. This network was fairly well characterized prior to the discovery of the association of PTP11 with Noonan Syndrome, so the network knowledge could be immediately applied to predict other nodes that, when perturbed, might also result in a Noonan Syndrome phenotype [13]. When tested, using an iterative network-refinement approach, several of these genes were also found to be fundamental to Noonan Syndrome etiology. These genes include RAF1, SOS1, HRAS, KRAS, and RAF1 [36,39,44,46,52].

Genetic association studies show great promise for breaking into etiologic network models for other human diseases. The technology for association studies for autoimmune diseases has been progressing steadily since the 1970s. However, until 2006, the only major genetic association with these diseases had been the MHC locus, which encodes the molecular complexes that train the immune system to recognize particular peptides as activators. The accumulated progress in technology moved research across a discovery threshold early in the twenty-first century and a large number of additional gene associations in a variety of autoimmune diseases have been discovered, as well in other complex human diseases. These include the discovery of IFIH1 for type 2 diabetes [45], ITPR3 for type 1 diabetes [40], and IL7R for multiple sclerosis [15,22]. Further progress in these diseases will be somewhat slower than for Noonan Syndrome, but there is every reason to believe that these larger and more complex networks can be modeled in clinically useful ways by applying the techniques discussed in the previous section.

In addition to genetic association studies, data analysis from high-throughput sources can also help enumerate key nodes in human networks. Carefully curated analyzes of orthologous high-throughput data sets can yield lists of nodes that are likely to be fundamental to predictive network models. One of the more promising human systems for this systems biology approach is modeling macrophage activation through innate immune receptors [2]. Non-predictive network diagrams based on extensive literature analysis are available for the macrophage [34]. Such diagrams facilitate the development of predictive network hypotheses [35].

Rather than breaking into an unknown network, Bergholdt et al. [7] have approached the type 1 diabetes network by assuming that it is a subnetwork of current iterations of the global human protein-protein interaction network. They take key-node analysis a step further by developing a statistic for key subnetworks. Significance is computed from genetic association data for all genes in the subnetwork, as well as the interactions between these genes. This approach is amenable to further refinement by the incorporation of additional orthologous data sets, such as function and expression, into the significance metric. Bergholdt et al. succeed in identifying a number of candidate subnetworks that may transduce signals that cause diabetes.

Human macrophages are tractable experimental systems even when taken out of the context of the human body, and much of the macrophage regulatory network operates in a near native fashion even when isolated from other human cells. Because systems biology of the human immune system can in cases be reduced to the analysis of single cells, it shares some simplicities with prokaryotic systems biology. Systems biology of the human immune system is therefore a natural initial focus for human systems biology.

A large number of high-throughput data sets have been acquired from macrophages, particularly time-series following stimulation through toll-like receptors. Several preliminary comparative network analyzes are available [21,37]. Reasonable near-term expectations from such studies, which may involve dozens of researchers and years of research, are the addition of one or several nodes and key interactions to a basic network such as the identification of the role of ATF3 as a important negative regulator of macrophage activation [14] or the bulk identification of many key nodes without their interactions such as the identification of all transcriptionally active transcription factors [33,41]. Studies of signaling networks in T-cells have also resulted in predictive models, but are impeded by the resistance of T-cells to some forms of genetic perturbations, such as RNAi interference [25].

A lot of clinical value is currently being generated through analyzes of high-throughput data sets that result in predictive models for disease [27]. In some cases these analyzes are done on only one type of data, such as microarray data, and are not orthologous data sets in the truest sense. Furthermore, the results are frequently statements of correlation, rather than causation, and thus are a fairly preliminary step towards a true understanding of the system, or a true predictive etiologic model. Nevertheless these predictions are amenable to iterative refinement with a good prospect that such refinements will eventually elevate these models to the level of understanding. Examples of such efforts include diagnosis and stratification.

Statistical methodologies such as discriminant or principle components analysis can extract variables that are highly informative for classifying systems. Quantitative values can be obtained from many thousand nodes in systems with known properties. For example, one can measure the expression of thousands of genes in a set of malignant tumors and controls. Simplified classifiers can be extracted from these many thousand variables. Clinically useful stratification of breast cancers into subtypes that respond preferentially to different therapeutic interventions has been one example of this approach [16,49,50]. These multivariate approaches also show promise for diagnosing disease, including acute infections [38], arthritis [4], and diabetes [24]. These diagnostic techniques show promise for detecting presymptomatic disease, in which cases the semantic line between prediction and diagnosis blurs considerably. These methodologies can also stratify disease, allowing for the tailoring of specific therapies to the disease. Additionally, responses to diseases, medications, therapies, or the environment can themselves be stratified, permitting the early identification of adverse reactions and changes towards a more positive therapeutic direction.

Future Directions

The most productive element of the systems biology paradigm is collaboration. Collaboration between scientists with similar skills allows projects of greater size and complexity to be tackled. Collaboration between scientists with different knowledge and skill sets permits knowledge to be combined in manners previously unimaginable and opens previously inaccessible vistas of knowledge. For decades, both types of collaboration have been widely used in other branches of science, such as meteorology, astronomy, and particle physics. Therefore the most promising future directions for system biology are those that maximize collaborations.

These collaborations will be enabled by new methods of data sharing, most likely leveraging the worldwide web. Interactive and updateable sources of information such as wikis and blogs will be key. Many online scientific journals have recently begun to encourage online commenting of papers. New methods of collaboration may evolve. These may involve virtual collaboration and might leverage peer-to-peer networks such as Facebook. The Nature publishing group has initiated such a tool, Nature Network.

New societal methods of career recognition and funding will likely stimulate systems biology. Traditional methods have recognized single-author or first author papers and tend to focus on journal publishing as a measure of productivity and utility. New methods will recognize collaboration and non-traditional measures of productivity, such as the creation of online resources. Recently, the NIH has altered its tenure track rules in recognition of these needs. Many of these changes will require a new generation of researchers willing to embrace them. Progress in systems biology will be increasingly driven by distributed worldwide collaborations that may involve researchers unaware of each other, with the collaborative element being contributions to virtual online projects. Projects will involve both specialists and generalists. In many cases data will be generated by a specialist in an experimental technique, and analyzed by distinct specialists in analytical techniques. Generalists will need to have skills in multiple disciplines and will be needed to direct projects and integrate data. Among the traditional twentieth-century fields of biology, immunology required extensive multidisciplinary knowledge from its practitioners; it is not surprising that many of the pioneers of systems biology were immunologists. Investigators trained in the clinical practice of medicine will be needed to set clinical priorities and recognize opportunities for the transfer of information from bench to bedside. Walls between disciplines will need to erode.

New and improved technologies will increasingly yield lower-costs, higher-throughputs, better sensitivity, better reproducibility and precision, better accuracy, and more comprehensivity [17,20]. Sequencing technologies continue to show promise for all of these characteristics, and dropping costs may permit them to underlie many measurements. Transcriptomics is currently performed mostly with arrays which may have fundamental upper limits on accuracy. Future transcriptomics may breach the accuracy barrier by employing transcript enumeration technologies that aim to sequence, for example, every mRNA in a single cell. Improvement in computational prediction and direct measurement of transcription factor binding to targets will permit these interactions to be confidently added to networks, and will vastly improve the predictive power of models of nuclear regulation. New methods of creating, storing, and sharing medical information electronically must be adopted that allow efficient use of aggregate information for research while preserving patient confidentiality and their rights to control their own information.

Effective improvements to human health cannot be made solely by a small group of individuals. Society must participate. Individuals must be informed, must support research, and must be willing to participate in and receive therapeutic interventions. Systems biology must therefore be alert to the broad needs of society [10]. These needs can help focus priorities. A society willing to modify its diet but reluctant to take pharmaceuticals would inform systems biology to prioritize models that predict the effects of altered diets over models of drug metabolism, and vice versa. Because human systems biology relies on human participation in research, such as in the Framingham Study and in gene association studies, efforts must be made to continue to educate all people in a deep inquiry-based scientific knowledge and to encourage them to share their personal medical histories and experience with the common knowledge pool. Current cultural shifts in the US towards increased guarding of personal information and mistrust of scientific information will create substantial challenges for the next generation of systems biologists.

Bibliography

Primary Literature

1. Aderem A (2007) Systems biology. Curr Opin Biotechnol 18(4):331–332
2. Aderem A, Hood L (2001) Immunology in the post-genomic era. Nat Immunol 2(5):373–5
3. Aderem A, Smith KD (2004) A systems approach to dissecting immunity and inflammation. Semin Immunol 16(1):55–67
4. Allantaz F, Chaussabel D, Stichweh D, Bennett L, Allman W, Mejias A, Ardura M, Chung W, Wise C, Palucka K, Ramilo O, Punaro M, Banchereau J, Pascual V (2007) Blood leukocyte microarrays to diagnose systemic onset juvenile idiopathic arthritis and follow the response to IL-1 blockade. J Exp Med 204(9):2131–44
5. Auffray C, Imbeaud S, Roux-Rouquie M, Hood L (2003) From functional genomics to systems biology: concepts and practices. C R Biol 326(10–11):879–92
6. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A (2005) Reverse engineering of regulatory networks in human B cells. Nat Genet 37(4):382–90
7. Bergholdt R, Størling ZM, Lage K, Karlberg EO, Olason PI, Aalund M, Nerup J, Brunak S, Workman CT, Pociot F (2007) Integrative analysis for finding genes and networks involved in diabetes and other complex diseases. Genome Biol 8(11):R253
8. Beyer A, Bandyopadhyay S, Ideker T (2007) Integrating physical and genetic maps: from genomes to interaction networks. Nat Rev Genet 8(9):699–710
9. Bornholdt S (2005) Systems biology. Less is more in modeling large genetic networks. Science 310(5747):449–51

10. Buchanan A, Califano A, Kahn J, McPherson E, Robertson J, Brody B (2002) Pharmacogenetics: ethical issues and policy options. Kennedy Inst Ethics J 12(1):1–15
11. Chaussabel D (2004) Biomedical literature mining: challenges and solutions in the 'omics' era. Am J Pharmacogenomics 4(6):383–93
12. Daly AK (2007) Individualized drug therapy. Curr Opin Drug Discov Devel 10(1):29–36
13. Gelb BD, Tartaglia M (2006) Noonan syndrome and related disorders: dysregulated RAS-mitogen activated protein kinase signal transduction. Hum Mol Genet 15, Spec No 2:R220–6
14. Gilchrist M, Thorsson V, Li B, Rust AG, Korb M, Roach JC, Kennedy K, Hai T, Bolouri H, Aderem A (2006) Systems biology approaches identify ATF3 as a negative regulator of Toll-like receptor 4. Nature 441(7090):173–8
15. Gregory SG, Schmidt S, Seth P, Oksenberg JR, Hart J, Prokop A, Caillier SJ, Ban M, Goris A, Barcellos LF, Lincoln R, McCauley JL, Sawcer SJ, Compston DA, Dubois B, Hauser SL, Garcia-Blanco MA, Pericak-Vance MA, Haines JL; for the Multiple Sclerosis Genetics Group (2007) Interleukin 7 receptor alpha chain (IL7R) shows allelic and functional association with multiple sclerosis. Nat Genet 39(9):1083–91
16. Hoadley KA, Weigman VJ, Fan C, Sawyer LR, He X, Troester MA, Sartor CI, Rieger-House T, Bernard PS, Carey LA, Perou CM (2007) EGFR associated expression profiles vary with breast tumor subtype. BMC Genomics 8(1):258
17. Hood L (2002) A personal view of molecular technology and how it has changed biology. J Proteome Res 1(5):399–409
18. Hood L, Galas D (2003) The digital code of DNA. Nature 421(6921):444–8
19. Hood L, Perlmutter RM (2004) The impact of systems approaches on biological problems in drug discovery. Nat Biotechnol 22(10):1215–7
20. Hood L, Heath JR, Phelps ME, Lin B (2004) Systems biology and new technologies enable predictive and preventative medicine. Science 306(5696):640–3
21. Iliev DB, Roach JC, Mackenzie S, Planas JV, Goetz FW (2005) Endotoxin recognition: In fish or not in fish? FEBS Lett 579(29):6519–6528
22. International Multiple Sclerosis Genetics Consortium, Hafler DA, Compston A, Sawcer S, Lander ES, Daly MJ, De Jager PL, de Bakker PI, Gabriel SB, Mirel DB, Ivinson AJ, Pericak-Vance MA, Gregory SG, Rioux JD, McCauley JL, Haines JL, Barcellos LF, Cree B, Oksenberg JR, Hauser SL (2007) Risk alleles for multiple sclerosis identified by a genomewide study. N Engl J Med 357(9):851–62
23. Janes KA, Yaffe MB (2006) Data-driven modelling of signal-transduction networks. Nat Rev Mol Cell Biol 7(11):820–8
24. Kaizer EC, Glaser CL, Chaussabel D, Banchereau J, Pascual V, White PC (2007) Gene expression in peripheral blood mononuclear cells from children with diabetes. J Clin Endocrinol Metab 92(9):3705–11
25. Kemp ML, Wille L, Lewis CL, Nicholson LB, Lauffenburger DA (2007) Quantitative network signal combinations downstream of TCR activation can predict IL-2 production response. J Immunol 178(8):4984–92
26. Kitano H (2002) Computational systems biology. Nature 420(6912):206–10
27. Kitano H (2007) The theory of biological robustness and its implication in cancer. Ernst Schering Res Found Workshop. (61):69–88
28. Kratz CP, Schubbert S, Bollag G, Niemeyer CM, Shannon KM, Zenker M (2006) Germline mutations in components of the Ras signaling pathway in Noonan syndrome and related disorders. Cell Cycle 5(15):1607–11
29. Lee ST, Ki CS, Lee HJ (2007) Mutation analysis of the genes involved in the Ras-mitogen-activated protein kinase (MAPK) pathway in Korean patients with Noonan syndrome. Clin Genet 72(2):150–5
30. Ma'ayan A, Jenkins SL, Neves S, Hasseldine A, Grace E, Dubin-Thaler B, Eungdamrong NJ, Weng G, Ram PT, Rice JJ, Kershenbaum A, Stolovitzky GA, Blitzer RD, Iyengar R (2005) Formation of regulatory patterns during signal propagation in a Mammalian cellular network. Science 309(5737):1078–83
31. Ma'ayan A, Jenkins SL, Goldfarb J, Iyengar R (2007) Network analysis of FDA approved drugs and their targets. Mt Sinai J Med 74(1):27–32. Ma'ayan A, Blitzer RD, Iyengar R (2005) Toward predictive models of mammalian cells. Annu Rev Biophys Biomol Struct 34:319–49
32. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 7 Suppl 1:S7
33. Natarajan M, Lin KM, Hsueh RC, Sternweis PC, Ranganathan R (2006) A global analysis of cross-talk in a mammalian cellular signalling network. Nat Cell Biol 8(6):571–80
34. Oda K, Kitano H (2006) A comprehensive map of the toll-like receptor signaling network. Mol Syst Biol 2:2006.0015
35. Oda K, Matsuoka Y, Funahashi A, Kitano H (2005) A comprehensive pathway map of epidermal growth factor receptor signaling. Mol Syst Biol 1:2005.0010
36. Pandit B, Sarkozy A, Pennacchio LA, Carta C, Oishi K, Martinelli S, Pogna EA, Schackwitz W, Ustaszewska A, Landstrom A, Bos JM, Ommen SR, Esposito G, Lepri F, Faul C, Mundel P, Lopez Siguero JP, Tenconi R, Selicorni A, Rossi C, Mazzanti L, Torrente I, Marino B, Digilio MC, Zampino G, Ackerman MJ, Dallapiccola B, Tartaglia M, Gelb BD (2007) Gain-of-function RAF1 mutations cause Noonan and LEOPARD syndromes with hypertrophic cardiomyopathy. Nat Genet 39(8):1007–12
37. Purcell MK, Smith KD, Hood L, Winton JR, Roach JC (2006) Conservation of toll-like receptor signaling pathways in teleost fish. Comp Biochem Physiol Part D Genomics Proteomics 1(1):77–88
38. Ramilo O, Allman W, Chung W, Mejias A, Ardura M, Glaser C, Wittkowski KM, Piqueras B, Banchereau J, Palucka AK, Chaussabel D (2007) Gene expression patterns in blood leukocytes discriminate patients with acute infections. Blood 109(5):2066–77
39. Razzaque MA, Nishizawa T, Komoike Y, Yagi H, Furutani M, Amo R, Kamisago M, Momma K, Katayama H, Nakagawa M, Fujiwara Y, Matsushima M, Mizuno K, Tokuyama M, Hirota H, Muneuchi J, Higashinakagawa T, Matsuoka R (2007) Germline gain-of-function mutations in RAF1 cause Noonan syndrome. Nat Genet 39(8):1013–7
40. Roach JC, Deutsch K, Li S, Siegel AF, Bekris LM, Einhaus DC, Sheridan CM, Glusman G, Hood L, Lernmark AA, Janer M on behalf of the Swedish Childhood Diabetes Study Group and the Diabetes Incidence in Sweden Study Group. (2006) Genetic mapping at 3-kilobase resolution reveals inositol 1,4,5-triphosphate receptor 3 as a risk factor for type 1 diabetes in Sweden. Am J Hum Genet 79(4):614–627

41. Roach JC, Smith KD, Strobe KL, Nissen SM, Haudenschild CD, Zhou D, Vasicek TJ, Held GA, Stolovitzky GA, Hood L, Aderem A (2007) Transcription factor expression in lipopolysaccharide-activated peripheral blood derived mononuclear cells. PNAS 104(41):16245–50
42. Roeder RG (2003) The eukaryotic transcriptional machinery: complexities and mechanisms unforeseen. Nat Med 9(10):1239–44
43. Schlitt T, Brazma A (2007) Current approaches to gene regulatory network modelling. BMC Bioinformatics. 8 Suppl 6:S9
44. Schubbert S, Zenker M, Rowe SL, Boll S, Klein C, Bollag G, van der Burgt I, Musante L, Kalscheuer V, Wehner LE, Nguyen H, West B, Zhang KY, Sistermans E, Rauch A, Niemeyer CM, Shannon K, Kratz CP (2006) Germline KRAS mutations cause Noonan syndrome. Nat Genet 38(3):331–6
45. Smyth DJ, Cooper JD, Bailey R, Field S, Burren O, Smink LJ, Guja C, Ionescu-Tirgoviste C, Widmer B, Dunger DB, Savage DA, Walker NM, Clayton DG, Todd JA (2006) A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. Nat Genet 38:617–9
46. Sovik O, Schubbert S, Houge G, Steine SJ, Norgard G, Engelsen B, Njolstad PR, Shannon K, Molven A (2007) De novo HRAS and KRAS mutations in two siblings with short stature and neuro-cardio-facio-cutaneous features. J Med Genet 44(7):e84
47. Tian Q, Stepaniants SB, Mao M, Weng L, Feetham MC, Doyle MJ, Yi EC, Dai H, Thorsson V, Eng J, Goodlett D, Berger JP, Gunter B, Linseley PS, Stoughton RB, Aebersold R, Collins SJ, Hanlon WA, Hood LE (2004) Integrated genomic and proteomic analyses of gene expression in Mammalian cells. Mol Cell Proteomics 3(10):960–9
48. Tijan R (1995) Molecular machines that control genes. Scientific American. 272(2):54–61
49. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415(6871):530–6
50. Weigelt B, Hu Z, He X, Livasy C, Carey LA, Ewend MG, Glas AM, Perou CM, Van't Veer LJ (2005) Molecular portraits and 70-gene prognosis signature are preserved throughout the metastatic process of breast cancer. Cancer Res 65(20):9155–8
51. Weston AD, Hood L (2004) Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. J Proteome Res 3(2):179–96
52. Zenker M, Horn D, Wieczorek D, Allanson J, Pauli S, van der Burgt I, Doerr HG, Gaspar H, Hofbeck M, Gillessen-Kaesbach G, Koch A, Meinecke P, Nowak A, Rauch A, Reif S, von Schnakenburg C, Seidel H, Wehner LE, Zweier C, Bauhuber S, Matejas V, Kratz CP, Thomas C, Kutsche K (2007) SOS1 is the second most common Noonan gene but plays no major role in cardio-facio-cutaneous syndrome. J Med Genet 44(10):651–6

Books and Reviews

Alon U (2006) An introduction to systems biology: Design principles of biological circuits. Chapman and Hall, Norwell

Bower JM, Bolouri H (2001) Computational modeling of genetic and biochemical networks. MIT Press, Cambridge

Capra F (1996) The web of life. Doubleday, New York

Carroll SB (2004) From DNA to diversity: Molecular genetics and the evolution of animal design. Blackwell Publishing Limited, Boston

Davidson EH (2006) The regulatory genome: Gene regulatory networks in development and evolution. Academic Press, San Diego

Fall CP et al (2002) Computational cell biology. Springer, New York

Goodsell DS (1992) The machinery of life. Springer, New York

Kaneko K (2006) Life: An introduction to complex systems biology. Springer, New York

Kauffman SA (1993) The origins of order: Self-organization and selection in evolution. Oxford University Press, Oxford

Kitano H (2001) Foundations of systems biology. MIT Press, Cambridge

Latchman DS (2004) Eukaryotic transcription factors. Elsevier, Amsterdam

Michal G (1999) Biochemical pathways. Wiley, New York

Palsson BO (2006) Systems biology: Properties of reconstructed networks. Cambridge University Press, Cambridge

Voit EO (2000) Computational analysis of biochemical systems: A practical guide for biochemists and molecular biologists. Cambridge University Press, Cambridge

Systems Biology, Introduction to

ROBERT A. MEYERS
Ramtech Limited, Larkspur, USA

Systems biology is a relatively new branch of life sciences that is highly interdisciplinary and aims to provide a quantitative analysis and understanding of life systems within a mathematical framework. Systems biology is the study of the interconnected aspect of molecular, cellular, tissue, whole animal and ecological processes, and comprises mathematical and mechanistic studies of dynamical, mesoscopic, open, spatiotemporally defined, nonlinear, complex systems that are far from thermodynamic equilibrium. The animating principle of systems biology is that the most important behaviors and networks of interacting elements (e. g.,molecules, cells, etc.), rather than by individual elements and interactions. This really necessitates the integrated application of high-throughput measurement technologies and advanced computational methods to model and predict biological responses. The 11 articles comprising this section describe the basic aspects and processes of systems biology and the relationship of systems biology to complexity and systems science. The basis and content of each of the 11 articles in this section are described below. There are six additional articles that were nominated and accepted into other sections which are connected to complexity in Systems Biology, and supplement the 11 articles in this section. These six are listed at the end of this Introduction.

With rare exceptions, all known living organisms encode their genetic material in the form of double-stranded DNA, in one or more chromosomes, collectively referred to as the "genome" (▶ Genome Organization). The genome includes most of the information needed by the cells to stay alive, to differentiate into new cell types, and to perform their functions in the context of the organism. As such, it is the ultimate resource for identifying the full set of components in the living system. Eukaryotic genomes are much larger than strictly needed to encode the relatively modest set of genes in them, but several mechanisms give rise to a very complex transcriptome. It is necessary to understand how the genome is organized and how it evolved and be able to build suitable null hypotheses for testing whether predictions are likely to be real, and thereby understand whether specific observations are likely to be biologically meaningful.

Sequence data alone is currently of limited use for identifying the functional elements of a genome and for elucidating how these elements interact to control physiological processes.

The field of functional genomics aims to meet these challenges using the sequence data as a blueprint (▶ Functional Genomics for Characterization of Genome Sequences). In a broad sense, functional genomics is defined as the large-scale experimental study of gene function and interactions. In the above article, the techniques and challenges in functional genomics are illustrated in the context of an ever more common case scenario: given a complete genome sequence, how does one figure out what the sequence means? The current state of the art is interpreted to arrive at the function or functions of a given genome sequence.

Among the most fundamental problems in biology is deciphering the relationship between genotype and phenotype in the complex system of life forms. Understanding a complex system broadly requires that one (1) identify the elements, (2) determine the function of each element, (3) identify and characterize the interactions between elements, and (4) assemble all of this information into a mathematical model that accurately simulates the system and predicts its responses to novel perturbations (▶ Systems Genetics and Complex Traits). In the context of systems biology, this process begins with sequencing an organism's genome and identifying the functional elements, e. g. the genes. There are numerous methods for determining the function(s) of a given gene product, where "function" can describe both the specific biochemical activity carried out by the gene product and the role that activity plays in the organism's response to an environmental or developmental stimulus. Understanding complex genetics is of increasing relevance to the study of human health and is essential to the development of predictive, preventive, and personalized medicine.

Systems biology is the derivation of emergent properties of a multicomponent biological system through the construction of quantitative predictive models. These predictive models are typically formulated as networks. Human systems biology is the practice of systems biology to elucidate emergent properties of humans or any subcomponent of the human body, such as a cell (▶ Systems Biology of Human Immunity and Disease). Human systems present challenges and opportunities not found in other organisms. Most problems in human systems biology are motivated by a desire to understand, predict, prevent, ameliorate, or cure a human disease. The human immune system is a mediator for most human diseases, and is the key system mediating autoimmune and infectious diseases.

Understanding the operation of cellular networks is probably one of the most challenging and intellectually exciting scientific fields today (▶ Biological Models of Molecular Network Dynamics). Cellular networks are some of the most complex natural systems we know. Even in a "simple" organism such as E. coli, there are at least four thousand genes with many thousands of interactions between molecules of many different sizes. With the availability of new experimental and theoretical techniques, our understanding of the operation of cellular networks has made great strides in the last few decades. An important outcome of this work is the development of predictive quantitative models. Such models of cellular function will have a profound impact on our ability to manipulate living systems which will lead to new opportunities for generating energy, mitigating our impact on the biosphere and last but not least, opening up new approaches and understanding of important disease states such as cancer and aging.

Biological research over the past century or so has been dominated by reductionism – identifying and characterizing individual biomolecules – and has enjoyed enormous success. Throughout this history, however, it has become increasingly clear that an individual biomolecule can rarely account for a discrete biological function on its own. A biological process is almost always the result of a complex interplay of relationships amongst biomolecules, and the treatment of these relationships as a graph is a natural and useful abstraction (▶ Biomolecular Network Structure and Function).

Broadly speaking, a biomolecular network is a graph representation of relationships (of which there are many types) amongst a group of biomolecules. Vertices or nodes represent biomolecules, including macromolecules such as genes, proteins, and RNAs, or small biomolecules like amino acids, sugars, and nucleic acids. An edge or link between two vertices indicates a relationship between the corresponding biomolecules, which could include physical interaction, genetic interaction, or a regulatory relationship (e. g., the protein product of gene A regulates the expression of gene B). This abstraction, although simplifying, converts a complex web of biological relationships into a mathematical graph, from which we can study its structural features as well as their implications on biological functions.

Data integration and model building have become essential activities in biological research as technological advancements continue to empower the measurement of biological data of increasing diversity and scale (▶ Biological Data Integration and Model Building). High-throughput technologies provide a wealth of global data sets (e. g. genomics, transcriptomics, proteomics, metabolomics), and the challenge becomes how to integrate this data to maximize the amount of useful biological information that can be extracted. Integrating biological data is important and challenging because of the nature of biology. Biological systems have evolved over the course of billions of years, and in that time biological mechanisms have become very diverse, with molecular machines of intricate detail. Thus, while there are certainly great general scientific principles to be distilled – such as the foundational evolutionary theory – much of biology is found in the details of these evolved systems. This emphasis on the details of systems and the history by which they came into being (i. e. evolution) are distinct features of biology as a science, and influence the need for large-scale data integration. Also, biological systems are responsive to varying environments, with potential system states influenced by the combinatorics of all possible molecular and environmental perturbations. Thus, data space in this realm is extraordinarily large. There is no shortage of possibilities to explore, and vast amounts of data will be needed to reengineer biological systems and understand their workings at a (near) complete level of detail or at a level where accurate prediction can be achieved.

Another reason data integration is essential, is that biology arises through the complex interworking of many components. Thus, to understand biology we must integrate biological data in a way that will allow us, not only to access all our acquired data efficiently, but even more importantly allow us to study, predict, modify, and even engineer the emergent properties of biological systems, where the whole is more than the sum of the parts.

Boolean Networks are a class of discrete dynamical systems that can be characterized by the interaction of a set of Boolean variables. Random Boolean Networks, which are ensembles of random network structures, are a simple model class for studying dynamical properties of gene regulatory networks (▶ Boolean Modeling of Biological Networks). Boolean Networks have been used as generic models for dynamics of complex systems of interacting entities, such as social and economic networks, neural networks, as well as gene and protein interaction networks. Despite their conceptual simplicity, Boolean Networks exhibit complex nonlinear behaviors that are, to this day, a challenging object of investigation for theoretical physicists and mathematicians. Further, a discretization of gene expression is often regarded as an experimentally justifiable simplification making Boolean network models attractive tools for the biologist.

Many processes in cell biology, such as those that carry out metabolism, the cell cycle, and various types of signaling, are comprised of biochemical reaction networks. It has proven useful to study these networks using computer simulations because they allow us to quantitatively investigate hypotheses about the networks. Deterministic simulations are sufficient to predict average behaviors at the population level, but they cannot address questions about noise, random switching between stable states of the system, or the behaviors of systems with very few molecules of key species. These topics are investigated with stochastic simulations (▶ Stochastic Models of Biological Processes). The dominant types of stochastic simulation methods that are used to investigate biochemical reaction networks are covered, as well as some of the results that have been found with them. As new biological experiments continue to reveal more detail about biological systems, and as computers continue to become more powerful, researchers will increasingly turn to simulation methods that can address stochastic and spatial details.

Systems biology has various definitions. Common features among accepted definitions generally involve the description and analysis of interacting biomolecular components. Systems analysis of a biological network is quickly demonstrating its utility as it helps to characterize biomolecular behavior that could not otherwise be produced by the individual components alone. Three areas in which systems analysis has been implemented in biology include: (1) the generation and statistical analysis of high-throughput data in an effort to catalog and characterize cellular components; (2) the construction and analysis of computational models for various biological systems (e. g., metabolism, signaling, and transcriptional regulation); and (3) the integration of the knowledge of parts and computational models to predict and engineer biological systems (synthetic biology). Metabolism, as a system, has played an important role in the development of systems biology, especially in the modeling sense (▶ Metabolic Systems Biology). This is because the network components (e. g., enzymes and metabolites) have been studied in detail for decades, and many links between components have been experimentally characterized. Metabolic systems biology, compared to systems biology in general, entails the computational analysis of these enzymes and metabolites and the metabolic pathways in which they participate. Metabolic systems biology, using genome-scale metabolic network reconstructions and their models, has helped (1) to elucidate biomolecular function; (2) to predict phenotypic behavior; (3) to discover new biological knowledge; and (4) to design experiments for engineering applications. Constraint-based methods have played a pivotal role in the analysis of large and genome-scale metabolic networks. The structure, mathematical formulation, and analytical techniques of constraints-based methods have also paved the way for the successful modeling of other complex biological networks, such as transcriptional regulation and signaling networks.

Finally, ecological systems are paradigmatic examples of complex systems. For example, consider the thousands of species interacting in complex ways within rich communities such as tropical rainforests or coral reefs. The most pressing questions ecologists face deal with concepts such as stability, resilience, thresholds and non-linearities which are at the core of the sciences of complexity. How robust are these cathedrals of biodiversity? At which rate will they disassemble as a consequence of global change. For example, one of the long-standing questions in ecology is the relationship between complexity and stability. A review of some of the applications of the complexity sciences in the realm of ecological systems is presented (▶ Ecological Systems). Predicting the consequences of global change on biodiversity requires an interdisciplinary approach in which complexity approaches may be very useful. Information theory and diversity, networks, complex dynamics and spatiotemporal dynamics are all discussed relative to ecology as a complex system.

Additional articles connected to Systems Biology:

▶ Consciousness and Complexity
▶ Molecular Evolution, Networks in
▶ Complexity in Systems Level Biology and Genetics: Statistical Perspectives
▶ Fractals in Biology
▶ Biochemistry, Chaotic Dynamics, Noise, and Fractal Space in
▶ Exobiology and Complexity

Systems and Control, Introduction to

MATTHIAS KAWSKI
Department of Mathematics and Statistics,
Arizona State University, Tempe, USA

Control of dynamical systems has a long history: Watt's automatic centrifugal governor designed to regulate steam engines in the 1780s is considered a precursor of modern feedback control. Similarly, Bernoulli's work in the 1690s on the brachystochrone problem is a progenitor of optimal control. A distinguishing external feature of *controlled* dynamical systems is the presence of inputs that interact with the system in the form of deliberate controls, or unavoidable perturbations. Commonly control systems also have outputs (observations) that typically provide only incomplete information about the state of the system. Systems and control theory utilizes a broad array of mathematical disciplines – but a uniting, characteristic feature is the kind of questions being asked. In the 1950s the field became more formalized, it began to develop its own distinctive identity, and it has rapidly evolved ever since. Naturally, the linear theory developed first and quickly became a mature subject, with controllers now pervading everyday life, from thermostats, dozens of controllers in every automobile, to cell phones, attitude control of satellites, electric power networks, highway traffic control, to name just a few.

The articles in this section focus on the more recent developments, articulating how the fundamental problems and questions of systems and control theory are addressed in ever new settings, on new tools, and new applications.

Whereas in classical dynamical systems one tries to predict the uniquely determined future, a basic distinguishing question in control theory asks whether it is possible to steer the system to any desired target. This question about *controllability* has a simple and elegant answer in the case of linear systems. While much also is known for nonlinear systems, many challenges remain in both the finite dimensional setting [see ▶ Finite Dimensional Controllability] and in the setting of *distributed systems* modeled by partial differential equations [see ▶ Control of Non-linear Partial Differential Equations].

Assuming controllability using controls that are functions of time, the next question asks whether one can *automate* the control using feedback, that is, by making it a function of the state or of a measured output. Whereas for linear systems the theory is straightforward, unavoidable topological obstacles make nonlinear feedback stabilization problems much more challenging [see ▶ Stability and Feedback Stabilization]. A notion dual to controllability is that of observability: Rather than asking about the map from inputs to the state, a system is observable if the history of the outputs (measurements) determines the state of the system [see ▶ Observability (Deterministic Systems) and Realization Theory]. The typical engineering problem involves much more complex multi-tasking, dealing with multiple objectives in the presence of diverse constraints. Much of this well developed theory of output regulation, disturbance rejection, while attending to performance criteria, has recently been generalized to nonlinear systems [see ▶ System Regulation and Design, Geometric and Algebraic Methods in]. Generally, if there is one, then there are many control strategies that achieve a certain task – naturally one likes to single out a strategy that is the best. This leads to optimal control theory, which supersedes much of the classical calculus of variations, and which has rich interfaces with Lagrangian and Hamiltonian mechanics [see ▶ Maximum Principle in Optimal Control].

In many applications all one has to work with are sampled data for inputs and measured outputs. Moreover, commonly these are also corrupted by measurement errors and noises. A fundamental problem is to devise algorithms to *identify* the best model system that could generate these input-output pairs. Characteristically such algorithm should work in real time and improve the model upon newly available measurements. This well developed broad field includes geometric approaches to stochastic systems [see ▶ Stochastic Noises, Observation, Identification and Realization with] as well as more formal learning theories [see ▶ Learning, System Identification, and Complexity].

Systems and control theory is unified by a common core of questions, but employs a diverse array of models. Consequently it interfaces a wide range of mathematical disciplines, utilizes many different mathematical tools, but also fosters new development in diverse subdisciplines. An area that has evolved closely together with control is nonsmooth analysis with its rich theory of tangent objects to nonsmooth sets. Not only are controlled dynamical systems naturally modeled by differential inclusions, but control intrinsically is full of objects that are not differentiable in a classical sense. Most important is the value function in optimal control which encodes the minimal cost (time) to the target. Closely related to classical dynamic programming, this function ought to satisfy the Hamilton–Jacobi–Bellmann equation, a first order partial differential equation, but the value function generally is not differentiable in a traditional sense [see ▶ Nonsmooth Analysis in Systems and Control Theory]. Complementary to the nonsmooth approach are differential geometric approaches which include Lie theory and symplectic geometry. Closely related is a functional analytic operator calculus introduced into control by Agrachëv and Gamkrelidze which linearizes problems by imbedding them into infinite dimensional settings. This dramatically facilitates formal manipulations of nonlinear objects, and also gives new ways of looking at the underlying geometry [see ▶ Chronological Calculus in Systems and Control Theory].

Many of the early successful engineering implementations of systems and control theory used models with both continuous time and states (differential equations). But with the advent of ubiquitous digital controllers as well as ever broadening fields of applications the theories for increasingly diverse models are becoming well developed, too. These include, systems with or mix of continuous and discrete states (suitable, for example, to model systems with hysteresis) [see ▶ Hybrid Control Systems] and discrete time Hamiltonian and Lagrangian systems [see ▶ Discrete Control Systems] There are many new applications with quickly developing theories such as control of biological and biomedical applications, social dynamics, manufacturing systems, financial markets, and many more. Some of the most intense research in recent years has focused on systems with distributed intelligent agents (controllers) each of which has access to only local information [see ▶ Robotic Networks, Distributed Algorithms for]. Compared to these emerging areas the geometric theory of mechanical systems is one of the most mature and thriving research areas with scores of exciting new problems. It serves as a role model for the depth of geometric insight and for its well-understood close interconnections with many other disciplines [see ▶ Mechanical Systems: Symmetries and Reduction].

Acknowledgment

This work was partially supported by the National Science Foundation through the grant DMS 05-09030.

Systems Genetics and Complex Traits

Gregory W. Carter, Aimée M. Dudley
Institute for Systems Biology, Seattle, USA

Article Outline

Glossary

Allele A copy or alternate version of a gene.

Complex phenotype A trait (or phenotype) caused by polymorphisms in multiple genes.

Diploid An organism with two copies of each chromosome.

Epistasis A genetic interaction in which the double-mutant phenotype is identical to one of the single-mutants (masking epistasis); sometimes used as a general term for genetic interaction.

Genetic interaction A relationship that characterizes how two (or more) genetic perturbations or alleles combine to affect a phenotype.

Genetic network A representation of functional relationships between genes, gene products, or gene perturbations, often displayed as a graph.

Genetic perturbation The modification of a gene's presence, structure, or activity by a change in the DNA sequence in or near a gene.

Genetic transformation Genetic modification of a cell by the introduction of exogenous genetic material.

Haploid An organism with one copy of each chromosome.

Haplotype A set of polymorphisms that are linked and thus inherited as a unit.

Homolog One of two or more genes with similar DNA sequence due to shared ancestry.

Hybrid A diploid offspring from the mating of individuals from two distinct populations.

Intercross The mating of individuals from two distinct populations.

Isogenic A term describing strains of model organisms that are derived from the same individual or inbred line, e. g. the yeast deletion collection, in which each strain differs only by the presence of a specifically engineered mutation (e. g. a gene knock-out).

Masking epistasis A genetic interaction in which the double-mutant phenotype is identical to one of the single-mutants.

Mendelian genetics The study of traits caused by variations of a single gene.

Microarray hybridization Sequence-specific binding of nucleic acid (DNA or RNA) to short nucleic acid sequences tethered to a microarray chip.

Molecular barcode A short, unique sequence inserted into DNA that can be used to identify or quantify a gene, gene product, or strain.

Network motif A small subnetwork repeated in a larger network, suggesting repeated organization for a specific function.

Phenotype An observable property of an organism; sometimes implies an aberrant property, e. g. the mutant has a phenotype.

Pleiotropy The phenomenon in which a mutation in a single gene causes multiple phenotypes.

Quantitative trait locus (QTL) A chromosomal region linked to a measurable phenotype.

RNA interference (RNAi) A method of silencing a gene's expression by introducing a small RNA molecule with sequences complementary to a portion of the gene's mRNA transcript.

Reciprocal-hemizygote A hybrid in which one copy of a gene or chromosomal region has been removed.

Sporulation The formation of spores in fungi, produced by meiosis.

Synthetic sickness or lethality (SSL) A genetic interaction in which two genetic perturbations without individual fitness effects combine to cause a fitness defect or inviability.

3′ Untranslated region (UTR) The RNA sequence that follows the protein-coding region of a messenger RNA.

Definition of the Subject

Among the most fundamental problems in biology is deciphering the relationship between genotype and phenotype. Genetics, the study of how the DNA sequence of an individual (genotype) affects an observable characteristic (phenotype), has many properties of a systems science. Organisms contain large number of protein-coding genes, with ∼500 predicted in the simple bacterium *Mycoplasma genitalium*, ∼25,000 in the human genome, and over 50,000 in some plants. The relationship between genotype and phenotype can be complex, with multiple traits per gene (pleiotropy) and multiple genes per trait (complex phenotypes). Genes are the blueprints for both RNA transcripts and proteins that interact physically and functionally to produce the emergent behaviors that characterize life. Furthermore, biological systems can be systematically studied by perturbing elements of the system, such as genes or environmental factors, and quantitatively measuring the effects. This research strategy is generally considered a central component of systems biology [39,40], and here we refer to its application to the genotype-phenotype problem as *systems genetics*.

Understanding a complex system broadly requires one to (1) identify the elements, (2) determine the function of each element, (3) identify and characterize the interactions between elements, and (4) assemble all of this information into a mathematical model that accurately simulates the system and predicts its responses to novel perturbations. In the context of biology, this process usually begins with sequencing an organism's genome and identifying the functional elements, e. g. the genes. There are numerous methods for determining the function(s) of a given gene product, where "function" can describe both the specific activity carried out by the gene product and the role that activity plays in the organism's response to an environmental or developmental stimulus. For example, molecular or biochemical methods may reveal a protein to be enzyme that converts a substrate to a product, while genetic analysis may reveal that this activity is required for growth in the presence if a specific drug. Similarly, there are both molecular and genetic means of determining which gene products interact with each other, by detecting physical interactions (e. g. protein binding) and genetic interactions (e. g. a mutation in one gene modifies the effect of a mutation in a second gene). These interactions vary with respect to the nature, strength, and directionality of the interaction. Integrating these data into a coherent computational model of the system is a major challenge of systems biology, with early examples including a comprehensive analysis of the galactose utilization pathway in yeast [40] and bacterial chemotaxis [70].

Systems genetics approached to this problem use large-scale, high-throughput methods to decipher the network of gene functions and genetic interactions in an organism. Apart from their genome-wide scale, the data and methods used often mimic those of conventional genetic analysis. In classical Mendelian genetics, genes are defined as the entities by which traits are inherited. A specific trait exhibited by an organism thus depends on the version(s) of the corresponding gene, known as an allele, carried by that individual. Modern molecular biology has narrowed the concept of a gene to a distinct DNA sequence that is transcribed to encode an RNA or protein product. However, few phenotypes are entirely determined by a single gene. In most cases, multiple genes interact to confer a phenotype. These interactions can vary from simple modifier relationships, in which an allele of one gene alters the phenotype of another gene, to multiple genes with complicated interdependencies. The study of how multiple genes interact to influence a trait is generally known as complex genetics.

Understanding complex genetics is of increasing relevance to the study of human health and is essential to the development of predictive, preventive, and personalized medicine. Many diseases, such as asthma, diabetes, heart disease, and cancer, show a degree of heritability that cannot be traced to a single gene [3]. Moreover, risks inherited from identified disease genes are often modified by background gene variation [11,22,33,34]. Genetic risk in these cases is determined by an individual's allelic profile across many genes, combined with multiple environmental factors. Many of these allelic variations consistently appear together, as haplotypes, which are being mapped in the human HapMap project [1]. Multiple genes involved in cancer are being cataloged in The Cancer Genome Atlas [17]. This data must be combined with new modeling techniques to develop models for understanding the multigenic, molecular basis of human diseases. Developing new treatments for diseases with genetic susceptibilities will require not only the ability to genotype and classify patients on the basis of molecular fingerprints in tissues, but also an understanding of how genetic variants interact to affect clinical outcomes.

Introduction

Genetic interaction refers to the phenotypic effect of combining two or more genes with allelic variations. Genetic interactions are ideally observed under constant environmental conditions and in isogenic strains of model organisms, so that the lone variables are the genetic perturbations (i. e. the allele forms) carried by the strains being compared. Since genetic variants often combine in complex ways to affect phenotypes, it is necessary to infer and model the functional interactions between trait genes rather than viewing each gene as an independent factor. These functional interactions include all modes of activation or repression of one gene's activity by another. Take as an example, a measurable phenotype and a wild-type strain that defines the genetic background. Next consider two genes, A and B, which can be varied from that of the wild-type strain by mutations such that the (isogenic) genetic background is otherwise unchanged. Taking cell growth in the presence of the drug caffeine as an example phenotype, imagine that a mutation in A inhibits growth in the presence of caffeine. The presence of an additional mutation (B) in the strain (i. e. an AB double mutant) could have one of many different effects. B could of suppress the phenotype of A (restoring wild-type growth on caffeine), enhance the phenotype (decreasing growth on caffeine further than that observed in the A single mutant), or show no effect at all (growth equal to that of the A mutant). Interactions such as these have historically been used to map functional pathways [36].

In addition to the direction of the effect, the magnitude of phenotypic suppression or enhancement also provides important information about the nature of the genetic interaction. Additive effects, where the magnitude of an AB double mutant is equal to the combined effect of each individual's phenotype, are consistent with independent (non-interacting) genes in the same genetic pathway (e. g. where A and B both confer a phenotype) or in different genetic pathways (e. g. where A displays the phenotype, but B does not). Synergistic effects describe an AB double mutant trait that is quantitatively more extreme than would be expected from the linear combination of both. Perhaps the best studied class of synergistic interaction is synthetic lethality [21], in which the combination of two non-lethal mutations, A and B, combine to confer lethality in the AB double mutant. The concept has been appreciated for many years in both fruit fly [75] and yeast [30], and synthetic lethal interactions have identified redundant functions in a common (essential) biological pathways and parallel functions in separate pathways that are able to buffer each other [32]. Such buffering against genetic defects has long been believed to confer a fitness advantage [84]. The buffering of disease-related genes is of particular interest in identifying alleles that either alleviate or aggravate effects within individuals across a genetically diverse population.

In addition to assigning genes to functional pathways and uncovering functional relationships between pathways, the direction and magnitude of genetic interactions can also be used to determine the fine-structure of a genetic network and information flow through the system. Analysis of synthetic and epistatic interactions (Fig. 1) can be used to determine functional topology and the direction of information flow, commonly referred to by geneticists as "ordering genes in a pathway" [4]. Taking again the example of a mutation in gene A that inhibits growth in the presence of caffeine, now consider a mutation in another gene, C, that enhances growth under that condition. If the phenotype of an AC double mutant is a caffeine growth defect equal to that of the A single mutant, A is said to be epistatic to C. Such an interaction is often interpreted as evidence that A acts downstream of C, in that its mutation masks any modifications that occur further upstream of the phenotypic output [4]. This hypothesis is an example of serial information flow through the system, with the perturbed genes influencing each other and the growth phenotype in a linear sequence.

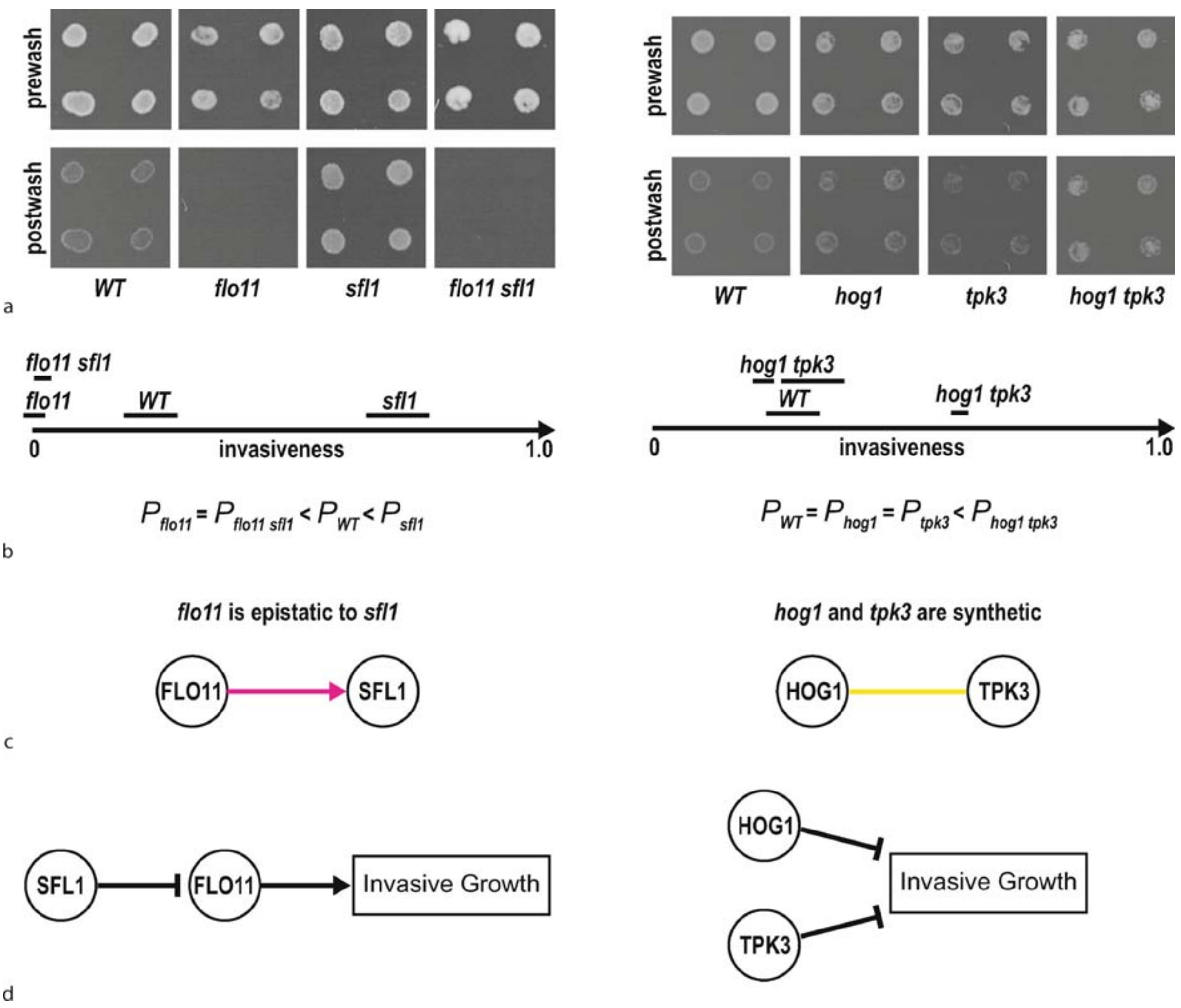
While genetic interaction analysis cannot, when taken alone, decipher the biochemical activity of the gene products, it can establish functional relationships between gene pairs. One can thus view a genetic interaction as providing a comprehensive view of the coordinated activity of two genes, encompassing every step in cellular processing between genotype and phenotype. A broad picture of complex genetic control emerges when these interactions are assembled into a network.

Measuring Genetic Interactions Genome-Wide

Since assaying a genetic interaction requires measuring the phenotype of interest in multiple strains, inferring genetic interactions on a genome-wide scale has historically been intractable. The application of high-throughput assays in model organisms with sequenced genomes has made the systematic study of complex genetics possible. To date, these studies have primarily involved the yeast *Saccharomyces cerevisiae*, which is easy to genetically modify and has a large community of researchers committed to the public dissemination of genome-wide reagent sets, and the nematode worm *Caenorhabditis elegans*, which has proven amenable to perturbations by RNAi. The large-scale study of how a genotype contributes to the control of a phenotype has been greatly aided by advances in methods to both systematically perturb genes and measure phenotypes on a genomic scale. These high-throughput methods produce large, quantitative data sets, requiring parallel development of computational and numerical modeling methods to interpret the output in terms of biological function. The primary objective of these methods is to identify functionally relevant genes and describe how they influence one another to generate cellular activity. Using a systems biology approach to integrate this knowledge with other data types has the potential to generate models capable of validating biological insights and identifying high-priority candidate molecules for targeted therapeutic intervention.

Synthetic Sickness and Lethality

Functional profiling of individual genes is most directly carried out by deleting the gene and observing phenotypic consequences. Through advances in genomics [87], a library of yeast mutants [27] now exists with deletions for every gene in the background of the laboratory strain derived from S288c [86], which was sequenced in the yeast genome project [29]. The yeast genome contains ~ 6000 protein coding genes (also know as open reading frames or ORFs). Approximately 5000 of these genes are nonessential and were deleted by replacing the gene of interest with a "marker gene" that confers drug resistance and two molecular barcodes, short DNA sequences that uniquely identify each deletion mutant. The remain-

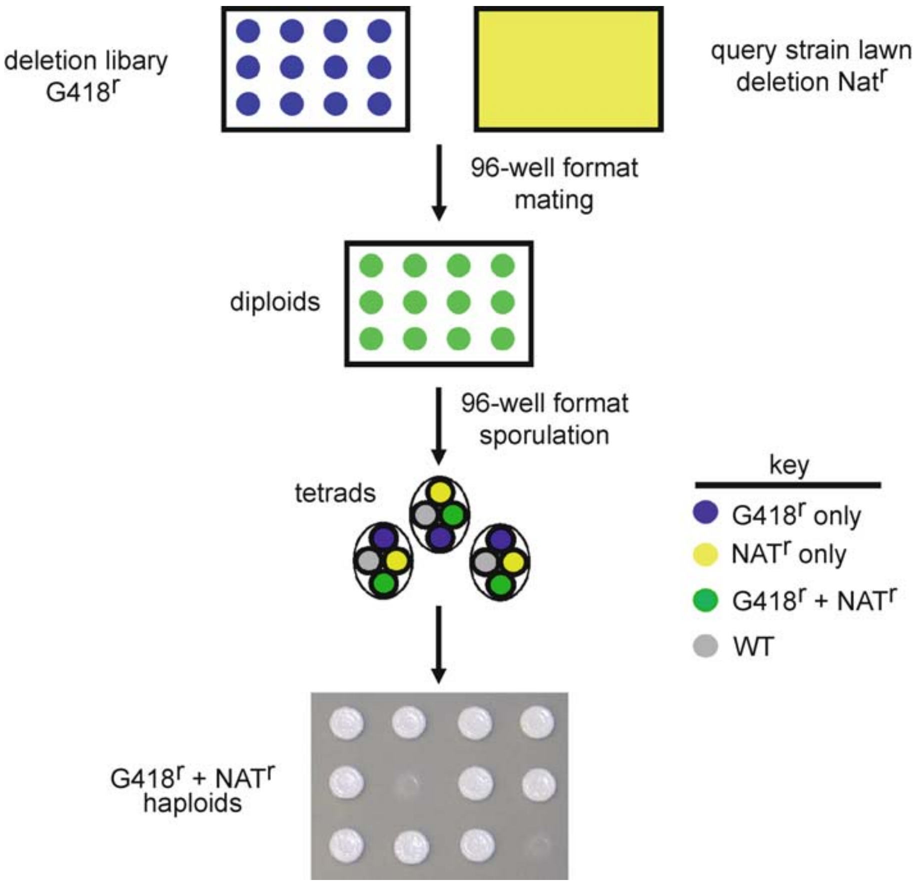

prewash
postwash
WT
flo11
sfl1
flo11 sfl1
a
hog1
tpk3
hog1 tpk3
flo11 sfl1
flo11
WT
sfl1
0
invasiveness
1.0
hog1 tpk3
WT
hog1 tpk3
$P_{flo11} = P_{flo11\ sfl1} < P_{WT} < P_{sfl1}$
$P_{WT} = P_{hog1} = P_{tpk3} < P_{hog1\ tpk3}$
b
flo11 is epistatic to sfl1
hog1 and tpk3 are synthetic
FLO11
SFL1
HOG1
TPK3
c
Invasive Growth
d

Systems Genetics and Complex Traits, Figure 1

Two examples of genetic interactions for yeast invasion. *Left* panels show the interaction between deletions of *SFL1* and *FLO11*, and the *right* panels show the interaction between deletions of *HOG1* and *TPK3*. a Yeast invasion assays for relevant strains. Yeast colonies (*light spots*) were grown for two days on nitrogen-limiting media (prewash) and cells were scraped from the surface (postwash). b Quantification of invasive growth, as the ratio of photo-assay signal before and after wash. Experimental ranges were determined by the variation among replicate colonies. c Genetic interactions as classified by Drees et al. [23]. d Classical genetics interpretations of pathway ordering. Epistasis (*left*) implies the genes are sequential in a pathway, and synthesis (*right*) implies the genes lie in parallel pathways. (Adapted from [23])

ing ∼ 1000 genes are essential for cell viability under standard laboratory conditions and thus must be maintained in strains that contain another functional copy of the gene. The lists of essential and non-essential genes are continually updated [44] as ORF annotation and verification improves [46] and closely-related strains of *S. cerevisiae* that have slight variations in essential genes are discovered [63]. Nonetheless, the observation that only 15–20% of yeast genes are essential, at least under optimal growth conditions, suggests that the genetic network governing yeast growth buffers the cell against genetic variation [32] and implies that multiple genetic mutations may compromise cell growth to a much greater degree than that observed for any single mutation.

Using the yeast deletion library as a starting point, three methods – synthetic genetic array (SGA), diploid-based synthetic lethality analysis with microarrays (dSLAM), and epistatic miniarray profiles (E-MAP) – have been developed as comprehensive, high-throughput assays for detecting pair-wise genetic interactions that



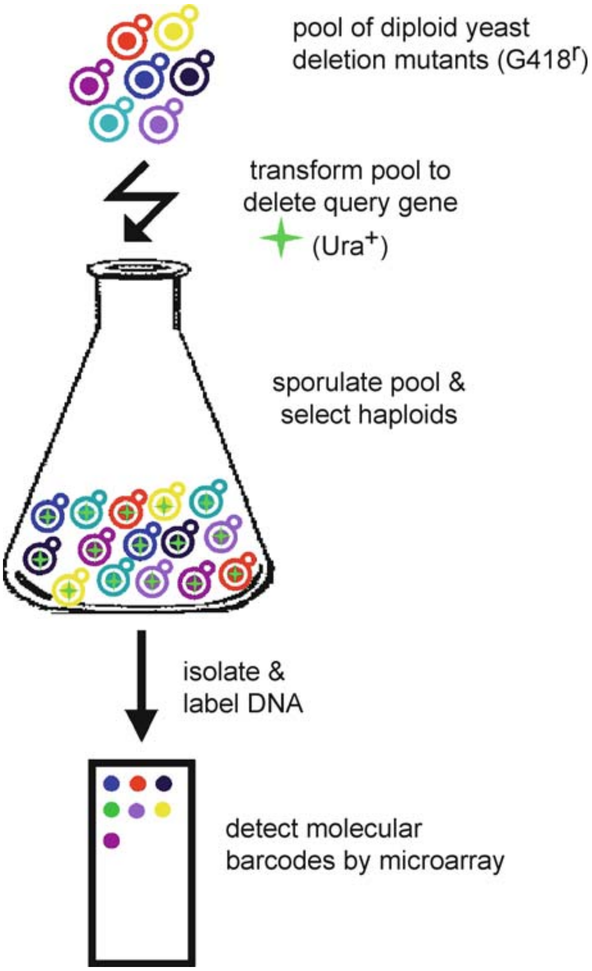
Systems Genetics and Complex Traits, Figure 2
Overview of the synthetic genetic array (SGA) and diploid-based synthetic lethality with microarrays (dSLAM) methods. Many of the experimental details have been simplified, more detailed descriptions are presented in [6]. In the SGA method a haploid strain (*yellow*) is constructed that contains a deletion of the "query gene" marked by a selectable drug marker (Nat^r) and a specialized reporter gene that facilitates high-throughput selection of haploid progeny that have mated and undergone sporulation 79. The query strain is mated to an ordered grid, haploid yeast deletion library (*blue*) marked by the selectable drug marker G418^r, forming diploids (*green*) that contain both deletion mutants, but which also contain a functional copy of each gene. Those diploids are then sporulated, allowing random shuffling of the genes and produces haploid progeny, which will only contain one copy of each gene of interest, i. e. the deleted or functional copy of the query gene. In the final step (photograph of actual yeast grids), haploid progeny that contain both the query gene (Nat^r) and the deletion strain present in that position of the grid (G418^r) are selected. The ability of the haploid progeny to grow in the presence of both drugs indicates that both deletions are present in the same strain, i. e. no interaction. Lack of growth indicates the inability to recover the double mutant combination, i. e. synthetic lethality

confer synthetic growth defects. Experimental details of these methods have been extensively reviewed [6,48,56]. We briefly summarize them here, as data generated by these methods dominates the genetic interaction data currently available. We then discuss other methods now being applied to comprehensively measure genetic interactions in yeast, worm, and mouse.

The synthetic genetic array (SGA) method (Fig. 2) was developed by Boone and colleagues [79,80] and extended to incorporate essential genes by transcriptional control via an artificial promoter [20]. The availability of a variety of drug-resistance markers enables a simple strategy for efficient construction of double gene deletion strains in yeast. Each strain in the haploid yeast deletion library harbors a gene (Kan) that confers resistance to the drug G418. In the SGA method a second haploid strain is constructed. This strain is a different mating type, and contains a deletion of the gene of interest marked by a gene (Nat) encoding resistance to a second drug (nourseothricin) and a reporter gene that enables selection for haploid progeny strains. This strain is mated to the deletion library, the resulting diploid strains are sporulated, and haploid progeny resistant to both drugs are selected. Synthetic lethality is indicated by the inability to recover haploid strains with both drug markers, suggesting that this combination of mutant alleles is inviable. Double mutants which show significant growth defects (so called "synthetic sickness") can also be identified by this method. The throughput of the method is increased by robotic manipulation of strains and phenotype scoring by image analysis software. To date, over 400,000 interactions have been tested and tens of thousands have been discovered [61], providing sufficient data for sophisticated network analysis and a valuable source of functional relationships between hundreds of yeast genes [80].

The issue of precise growth quantification of double-mutants has been addressed by a modification of the SGA method. In this approach, called epistatic miniarray profiles (E-MAP), growth rates are quantified and genetic interactions are identified as cases in which a double mutant strain exhibits a growth rate significantly above or below the expectation based on the two single mutants [19,64]. Essential genes were tested in this study by decreasing mRNA abundance via destabilization of the 3' UTR. Precise quantification of growth rates was performed by photo-assays of colony size and, based on the assumption that genetic interactions are rare, identifying the most significant deviations from the expectation that a double-mutant growth rate will be the product of the two single-mutant growth rates. Interactions with expected growth rates were classified as neutral, those with growth rates above expectation were labeled alleviating, and those with growth rates below expectation were labeled aggravating. To date, pair-wise interactions between 424 genes that localized to the endoplasmic reticulum [64] and 743 *Saccharomyces cerevisiae* genes involved in various aspects of chromosome biology (e. g. DNA replication/repair, chromosome segregation and transcriptional regulation) [18] have been tested.

Diploid-based synthetic lethality analysis with microarrays (dSLAM) (Fig. 3) is an alternate, transformation-based approach to assaying synthetic lethal interactions [57,58]. The method relies on molecular barcodes that uniquely mark each deletion strain to identify double-mutant fitness defects by DNA microarray hybridization. Heterozygous diploid knockouts [27] are initially transformed with a reporter gene that enables their conversion to haploid mutants following sporulation. Then, the gene of interest is deleted via high-efficiency integrative transformation, and the pool of yeast strains is sporulated and separated into single and double mutants using the drug resistance markers. Genomic DNA containing identifying bar codes from each pool is isolated. Bar codes are am-


pool of diploid yeast deletion mutants (G418^r)
transform pool to delete query gene
(Ura$^+$)
sporulate pool & select haploids
isolate & label DNA
detect molecular barcodes by microarray

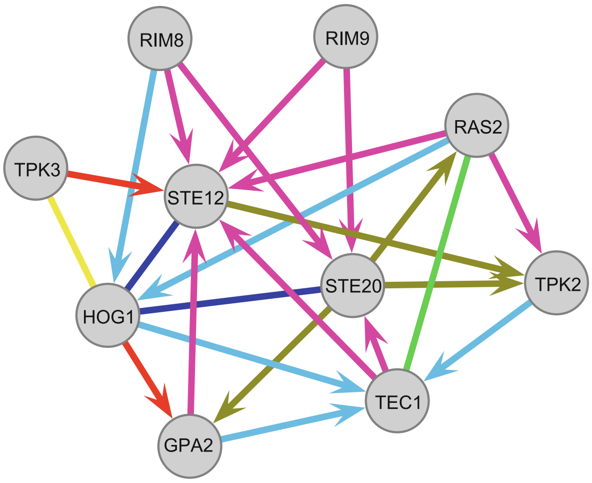
Systems Genetics and Complex Traits, Figure 3
The dSLAM method takes advantage of the molecular barcodes present in each deletion strain to facilitate the manipulation of strains in a pooled format. This method uses a diploid version of the yeast deletion set (heterozygous diploids) in which each strain contains both the gene deletion (G418^r) and a functional copy of the same gene. Prior to the analysis, the same reporter gene that facilitates high-throughput selection of haploid progeny used in SGA is introduced into all of the strains (not shown). The first step is the deletion of the query gene in (ideally) all the strains in the pool by transforming in a construct that will delete the gene of interest and contains a marker that will allow successfully transformed cells to grow in the absence of the nutrient uracil (Ura$^+$). The pool is then sporulated and haploid progeny containing both the G418^r deletion from the collection and Ura$^+$ query deletion are selected. DNA is then isolated from the pool of strains and the sequences containing the molecular barcodes are fluorescently labeled and hybridized to a DNA microarray that can detect the barcode sequences. Thus, microarray intensity provides a quantitative readout of the amount of each mutant recovered in the selection, with extremely low microarray intensity indicating synthetic lethality

plified, labeled, and hybridized to microarrays to quantify the ratio of single to double mutant growth rates. SSL interactions, also referred to as synthetic fitness or lethality (SFL) defects, are identified by high ratios that correspond to reduced double-mutant growth. Initial comparisons provide evidence that the dSLAM method achieves higher data consistency and lower false-negative rates than the SGA approach. dSLAM can also be modified to study dosage dependent SSL, genetic suppression, and haplo-insufficiency. Although the microarray based method potentially allows a more quantitative measure of synthetic sickness, the low signal intensity observed for many barcodes has hindered precise quantification. Because it relies on growth in mutant pools, the dSLAM method may be the most efficient method for genome-wide SSL screens.

The relatively recent introduction of RNA interference (RNAi) technology for gene knock-down allows multiple loss-of-function genetic perturbations to be carried out at high throughput [43,50] in model organisms such as the worm *C. elegans* and the fruit fly *D. melanogaster*. Large-scale RNAi based synthetic lethal screens have been performed in *C. elegans* [49]. Approximately 350 synthetic interactions between 160 genes in the worm EGF/Ras, Notch, and Wnt signaling pathways were discovered among 65,000 tested gene pairs. The highly-connected "hub" genes encode chromatin regulation proteins that are conserved across multiple species, which were hypothesized to be general buffers of genetic variation. Another study used the same experimental techniques to test functional redundancy between duplicated genes in *C. elegans* [78]. Synthetic interactions were detected in eleven percent of the duplicated pairs tested (16/143), suggesting that in some cases duplicated genes have a degree of functional redundancy. Interestingly, the majority of the interacting gene pairs were found to be duplicated over 80 million years ago, strongly suggesting that this functional redundancy can be conserved by positive selection.

Genetic Interactions for Quantitative Phenotypes

The vast majority of genetic interactions detected by the large-scale methods described above have been limited to the phenotype space of growth in rich yeast medium. However, phenotypes can be derived from any observable trait, including growth under a variety of environmental conditions, cell morphology, metabolite production, or gene expression pattern. To date, the most comprehensive analysis of genetic interactions detected using a phenotype other than growth in rich media include an analysis of interactions in nitrogen-poor media [23], or in the presence of a DNA damaging agent [71].



Systems Genetics and Complex Traits, Figure 4
Subsection of the yeast invasion network showing genes from various pathways and the variety of genetic interactions observed. Edge colors correspond to different interaction modes, e. g. *violet* is epistasis, *yellow* is synthesis, and *dark blue* is additivity. Subset of data from [23]

A mixture of gene deletions and high-copy, gain-of-function gene mutations were combined to map interactions in the yeast invasion network for nitrogen-poor media (Fig. 4) [23]. When starved for nitrogen, many strains of *S. cerevisiae* undergo cell differentiation into a pathogen-like, filamentous growth form and colonies subsequently invade solid media [28]. Drees and co-workers measured invasive growth by photographing colonies before and after vigorously washing cells from the agar surface (Fig. 1a). Multiple replicates of each genotype were assessed using image analysis software to quantitatively determine a mean invasion score and its variance (Fig. 1b). Since the phenotypes of the original single mutants ranged from non-invasive to strongly hyper-invasive, a wide variety of genetic interactions were possible. These interactions were first translated into mathematical inequalities that defined the relationship between the wild-type strain, the two single mutants, and the double mutant in terms of "less than", "greater than", and "equal to" relationships. For example, a synthetic interaction in which a double mutant (AB) exhibits a phenotype not shown by either single mutant (A or B) would be either "$AB < A = B = WT$" or "$A = B = WT < AB$". After systematically enumerating all 45 possible inequalities, the equations were grouped into nine "modes" of genetic interaction. Many of these interaction modes, such as synthesis, epistasis, or suppression, are well known to geneticists, but less familiar modes also appear. For example, the mode labeled "asynthesis" was defined as a double mutant phenotype which matched that of both single mutants but was distinct from the wild type. This general analysis was applied to over 130 genes and over 1800 pair-wise interactions were tested, revealing a rich spectrum of functional interactions that included examples of all nine modes.

A similarly multi-modal genetic interaction network was more recently constructed for genes that impart resistance to the DNA-damaging agent methyl methanesulfonate (MMS) [71]. Twenty-six such gene deletions (out of over 4700 tested) were identified by reduced growth rate in the presence of MMS, and double-mutant strains were constructed and assayed for all 325 pair-wise combinations. Although the selection method precluded the detection of synthetic interactions, this study is a quantitative analysis of genetic interactions among a set of genes implicated in a specific biological process.

Analysis of Complex Phenotypes that Result from Natural Variation

Although the majority of systems biology approaches to complex genetics have involved engineered genetic perturbations, genetic interaction studies are being extended to models of natural populations. A series of papers by Kruglyak and colleagues have attempted to infer genetic interactions in recombinant progeny strains [7,8,74]. In this work, two parent yeast strains, the sequenced "laboratory strain" (BY) and a "wild strain" isolated from a vineyard (RM) were mated and 112 haploid progeny strains were isolated. Each strain was genotyped at approximately 3000 markers and RNA expression levels were measured using DNA microarrays [9]. Statistical linkages were found between expression variations and quantitative trait loci (QTLs), and these loci were then assessed for interactions [74]. Although the data suggest that most transcripts were influenced by multiple loci [7], statistically significant pair-wise interactions have only been found for 225 transcripts [8]. Another series of yeast studies, taking high-temperature growth (HTG) as a quantitative phenotype, used reciprocal-hemizygosity as a tool to dissect QTLs in yeast [72]. The contribution of each allele to HTG was inferred by constructing hybrids of two yeast strains and then systematically deleting one copy of each gene within a candidate QTL. The QTL architecture was found to be more complex than expected, featuring both *in cis* and *in trans* linkages and suggesting that QTL regions cannot always be narrowed to a single local gene. The method was also used to explore how the influences of three trait genes were affected by variation in genetic backgrounds [69]. Reciprocal-hemizygosity analysis with ten other yeast strains demonstrated that the trait genes identified in one background did not necessarily contribute in the same way in other genetic backgrounds. These results demonstrate that, in addition to complexity within a QTL region, there is additional complexity due to genetic interactions. Together these studies illustrate some of the substantial complications that will arise in the analysis of complex genetics in the context of natural populations, such as the identification of trait loci in a background of genetic random variance, the separation of phenotypic effects from *cis*-acting and *trans*-acting polymorphisms, and the limitations of the pair-wise interaction analysis to complex genetics.

Furthermore, the logic of genetic interactions could be applied to functional interactions between different types of biomolecules. For example, recent work has studied the genetic basis of drug response in yeast [59]. Multiple polymorphisms were found that linked genetic loci to certain small molecule drugs, constituting a set of 124 gene-drug interactions. This approach provides a model for understanding variation in drug response within a genetically diverse population. Another study directly applied genetic interaction techniques to pair-wise interactions between drugs by mapping a set of compounds into classes that correspond to the cellular functions they affect [93]. These works begin to address how the understanding of complex genetics can be extended to involve additional system elements and, ultimately, generate a view of the organism as a complex system.

Overall, the study of genetic interactions is a field that has seen substantial development in the past decade. High-throughput methods will continue to open new avenues of research and expand established approaches to genome-wide scales. Data generated by these experiments will require formal and systematic approaches for analysis. Currently, network methods are the dominant framework.

Constructing Genetic Interaction Networks

Prior to high-throughput experimentation, genetic studies mapped interactions between a few genes in isolated functional pathways. These pathways represent information flow through a small number of elements that regulate a specific cell behavior, such as response to osmotic stress. Today, high-throughput techniques enable the assembly of large-scale genetic interaction data into networks and hold the potential for constructing semi-global maps of multiple, inter-connected pathways.

This wealth of data requires a formalism to efficiently analyze relationships between system elements. Complex systems are often best visualized as networks in which individual points, called *nodes*, are connected by lines or arrows, called *edges*. In genetic interaction networks, the nodes commonly represent genes, alleles, or gene perturbations (e. g. a gene deletion). Edges that connect nodes can take on different meanings, ranging from classic regulatory influences of one gene's activation or repression of another's activity to edges that signify a direct observation, such as synthetic lethality. The central challenge is to represent the richness of complex genetics in an informative way. Because the construction of each study is informed by a different biological problem, diverse network approaches have been pursued in the literature.

Since they are binary, synthetic lethal interactions are easily represented as a network of gene deletions connected by SSL edges. Although the formal information content of the edges is low, SSL networks have the advantage of relatively high coverage. Pair-wise combination of about 1000 yeast genes have been tested, and although this represents only 3% of all possible yeast pairs it encompasses 500,000 experiments. Four thousand of these pairs were SSL positive, producing a sparse but topologically interesting network. Non-essential genes averaged 34 synthetic lethal partners, and essential genes were found to have more interaction partners on average [80]. Extrapolating these results genome-wide, the authors estimate 200,000 essential pair-wise combinations in yeast, a number of lethal outcomes that is much larger than the 1000 genes that are essential individually [6]. The current network of 4000 genetic interactions demonstrates that SSL interactions are more likely to occur between non-essential genes of the same or related function, unlike SSL partners of essential genes that tend to show a broad spectrum of functions. This suggests that network location can be a predictor of gene function for non-essential genes, enabling novel functional hypotheses for genes based on a common function of interaction partners.

The known SSL network exhibits two interesting topological properties. First, it appears to be a *small world* network, defined as a network in which the length of the shortest path between two nodes is smaller than random expectation. Nodes are densely grouped into local clusters in which the interaction partners of any gene also tend to interact. Thus, the most likely candidates for SSL interactions for a given gene are the interaction partners of its known SSL partners, greatly increasing the likelihood of finding additional interaction partners in a targeted screen [80]. The second network property observed is that the number of SSL partners for each gene follows a power-law or *scale-free* distribution in both the yeast [80] and worm [49] data. As a result, most genes have relatively few interactions, and a small number of genes have relatively large number of interactions. This latter set, referred to as hubs, can be viewed as more important for fitness since the organism is most sensitive to pair-wise perturbations involving them. If this property is also a feature of human networks, the hub concept may inform the treatment of disease. Taking cancer as an example, hubs of cancer cell networks might be targeted for perturbation to increase the chance of a SSL interaction with a gene already genetically perturbed in the cancerous cell.

Nearly 5000 additional, and mostly novel, synthetic fitness or SSL interactions were discovered by studying genes involved in processes related to DNA integrity in yeast using dSLAM [57]. A total of 16 gene modules were identified in the network by clustering genes with a high congruence in SSL interactions, and each module was assigned a function based on its member genes. The resulting network mapped compensatory pathways among multiple biological processes, including DNA replication, DNA repair, checkpoint signaling, chromatin structure maintenance, and the response to oxidative stress. The modular architecture was consistent with prior models, and involved many genes with human homologs linked to cancer and aging.

The study of SSL interaction networks has the practical advantages of a simple, rapid assay and the direct mapping of each positive result to an edge in the interaction network. This narrow definition of genetic interaction, however, cannot explore other types of genetic interaction that have proved informative, such as suppression and masking epistasis. Encompassing all types of genetic interaction requires a systematic formalism such as the inequality relationships developed by Drees et al. [23]. As described above, these inequalities can be grouped into rules of genetic interaction corresponding to familiar and newly defined interaction modes. While many of these modes, such as synthesis, were symmetric with respect to the interacting genes, four of the modes produce directional edges. For example, instances of suppression were rendered with a colored arrow from the suppressor gene to its target gene. The resulting 1800 genetic interactions between 133 genes compose a genetic interaction network, derived with the *PhenotypeGenetics* plug-in for the *Cytoscape* software platform [67]. The nodes of this network included both gene deletions and plasmid-borne over-expressers, and the edges were mapped using nine colors corresponding to the nine modes of genetic interaction based on classical genetic interaction rules. This produced a richly multichromatic network suitable for informatic analysis (Fig. 4).

The multiplicity of genetic interaction modes facilitates the inference of relationships between individual genes and specific biological functions. In the Drees study, individual alleles were often found to interact in a particular mode, or *monochromatically*, with partner genes of coherent biological function, leading to hypotheses for regulatory and pathway organization [23]. Large-scale patterns of mutual information were also extracted from the data set, and groups of genes with significant mutual information between them formed network cliques corresponding to physical pathways. The genetic interaction patterns define a map of information flow from specific genetic perturbations to quantifiable phenotype effects.

This network was further analyzed by Taylor and Galitski [76], who searched for repeated subnetworks or *motifs* of three or four nodes, in which repetition suggested functional organization. Statistically enriched motifs were then assessed for functional coherence and many instances of functional monochromaticity were observed involving multiple genes. For example, deletions of multiple protein kinase genes were found to suppress gain-of-function perturbations of both signaling and transcription factor genes. This suggests that the kinases mediate information flow from the signaling proteins to transcriptional factors, and these pathways are broken by kinase deletions.

Multi-modal genetic interaction analysis was also performed on the MMS-sensitivity network of St. Onge et al. [71]. Genetic interactions were initially classified as aggravating, alleviating, or neutral interactions based on the double-mutant growth rates relative to that of the two single mutants. Alleviating interactions were of particular interest since there are many ways genes can mask or mimic another's effect. Thus, these interactions were sub-classified into five rules of interaction based on models of pathway ordering and protein co-function, four of which were assigned directionality. The resulting network model uncovered functional relationships and pathway order. The principle of genetic congruence, that genes with similar functional roles will share interaction patterns with functionally related genes in the system, was used to implicate a new regulator in DNA repair.

Both the yeast invasion and MMS networks classified genetic interactions into multiple modes based on interaction inequalities as described above. However, the authors used slightly different classification schemes. For example, in the invasion network "epistasis" included any interaction in which one single-mutant phenotype masks the phenotype of another. In the MMS network a further distinction was drawn between "masking epistasis", in which the single-mutant with a stronger effect masks the other, and "suppression", in which a weakly defective single-mutant masks the strongly defective one. While both networks proved informative, it is unclear which choice of mode classifications are the most meaningful for a given data set. This issue can be analyzed from the perspective of information theory. A measure of network complexity that places information in the context of the system under study could be used for the unsupervised classification of interaction modes for these two data sets. Network classification schemes that produce greater complexity might also generate more biologically meaningful information, i. e. more "biological statements" linking genes to functions were obtained in the more complex networks. Such complexity-based methods represent a powerful and general approach to genetic interaction analysis, with potential for the study of mammalian systems in which interactions are complex and gene annotation data is sparse.

Fully quantitative studies of genetic interactions have been hindered by the limited amount of numeric phenotype data required for more sophisticated mathematical techniques that might decipher the principles underlying genetic interaction. To overcome this limitation many researchers have adopted gene expression microarrays as a quantitative phenotype [9,41,42,47,62,85]. For each strain and environmental condition, genome-wide transcript levels constitute thousands of quantitative phenotype measurements that provide much greater molecular detail than a whole-cell phenotype, such as growth rate. This approach was combined with classical epistasis analysis by Van Driessche et al. [82]. By clustering mutants based on genome-wide expression profiles and applying the principle of masking epistasis, the authors accurately reconstructed the protein kinase A signaling pathway. In a more recent study, gene expression data was combined with whole-cell phenotype data and mathematical modeling to infer and predict genetic interactions relevant for yeast cell differentiation [12]. This analysis used combinatorial perturbations to infer how a set of regulatory genes quantitatively affect each other's activity as well as the expression of thousands of downstream genes. This quantitative model of genetic interaction in terms of genetic influences allowed the prediction of the effects of novel pair-wise genetic perturbations. The decreasing costs of microarrays and advances in genome-wide data analysis make gene expression data an increasingly attractive option for genetic interaction analysis.

Interpreting Genetic Interactions

The functional relationships described by genetic interaction networks reveal pathway organization and gene function. For predictive and therapeutic purposes, however, it is useful to translate these functional relationships into biomolecular hypotheses. Such hypotheses can suggest additional genetic perturbations to test and validate models, establishing the applicability of network analysis techniques and enabling an iterative process of model refinement. Moreover, the implication of specific alleles, molecules, and interactions in health-related processes can identify candidate genes and gene products for targeted therapeutic intervention with drug compounds or RNA interference. Thus, the interpretation of genetic interaction networks is necessary to improve the basic understanding of complex genetics as well as facilitate the development of personalized and predictive medicine.

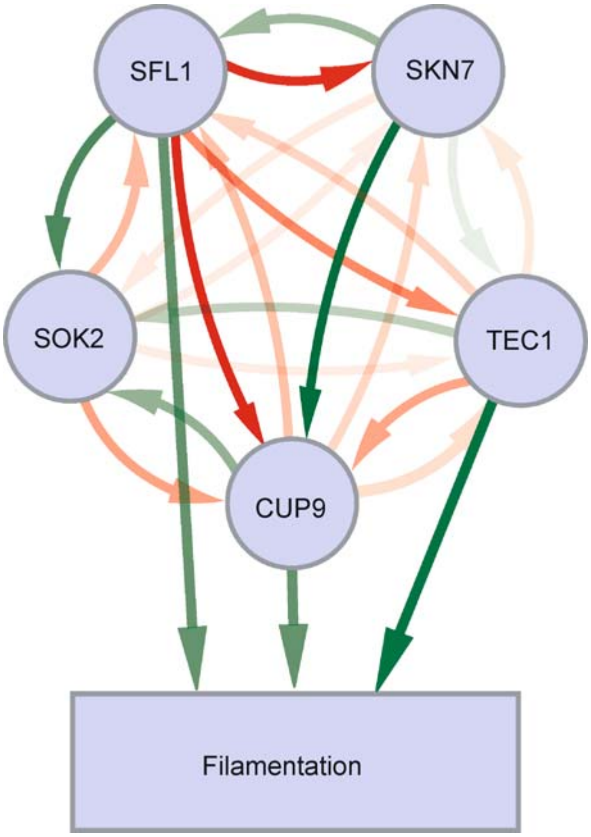
Formulating such hypotheses generally involves both molecular data integration and computational modeling. While genetic interactions define functional relationships and uncover pathways of information flow, molecular interaction databases provide a wiring diagram of possible mechanisms for the transmission of functional information. These complementary data types can be integrated to generate models of phenotype regulation [13,31,83,88]. Various computational methods enable this data integration. Kinetic modeling, for example, uses differential equations to predict the functional outcomes when a subset of physical interactions is active. Constraint based approaches use data analysis techniques to narrow the possible mechanisms that may have generated an observation, until the best hypothesis is identified [5,65,89,91]. Finally, whole-cell phenotypes can be dissected at the molecular level by integrating RNA expression levels and proteomic data [12,35,40]. These systems biology approaches have recently been applied to interpret complex genetic networks.

One particularly model-driven approach involves the use of computational models that simulate cellular processes to predict genetic interactions *in silico*. This requires a model, often highly detailed, that captures the relevant activity of every gene product involved in the process being simulated. Metabolic models based on Flux-Balance Analysis (FBA) have been successful in predicting growth rates for yeast [26] and bacteria [25] under a range of environmental conditions and genetic perturbations. Segré and collaborators used an FBA model to explore genetic interactions among metabolic genes in yeast by performing single and double gene knockouts *in silico* [66]. The model predicted that genetic interactions for growth rates fall into three distinct modes: cases in which the double deletion effect is significantly less than the product of the two single mutants (alleviating); cases in which the double deletion effect is significantly greater than the product of the two single mutants (aggravating); and cases in which the double deletion is not significantly different than the product of the two single mutants (neutral). The study also predicted similarity in interaction modes, or monochromaticity, between co-functional genes, a result that echoes the finding that genes with common functions preferentially interact in SSL networks. This work illustrates the power of using a system-wide computational simulation to predict complex genetic interactions using sensitive, quantitative whole-cell phenotypes. One of the major disadvantages of such an approach is the extensive amount of information necessary to formulate an FBA model. Furthermore, these predictions have not been empirically verified, and this would involve a major experimental undertaking.

As an alternative to model-based approaches, data-driven computational analysis has also been used to predict interacting gene pairs. Multiple types of relationships between gene products, such as common localization, function, and patterns of gene expression, were used to create a decision tree for predicting SSL interactions in yeast [88]. Genetic interactions were also predicted on a genome-wide scale in the nematode worm *Caenorhabditis elegans* by computationally integrating physical interactions, genetic interaction, gene expression, phenotype, and functional annotation data from yeast, worm, and fly [95]. The authors of this study used logistic regression to determine the predictive power of each type of data, or *feature*, based on a training set of 1816 known genetic interactions in *C. elegans*. Each feature was thereby weighted according to its predictive importance and the probability of interaction between every pair of *C. elegans* genes was calculated. In total, this procedure generated over 18,000 likely genetic interactions. As with previous studies, co-functional genes proved more likely to exhibit genetic interactions (although this is not an entirely independent result given that known genetic interactions between co-functional homologs were the predictors with greatest weight in the logistic regression). A handful of interaction candidates were experimentally tested using RNAi, and about half of the pair-wise perturbations exhibited some degree of genetic interaction in either a vulval development or pharyngeal pumping phenotypes. Interestingly, the vast majority of newly implicated genes did not have vulval development phenotypes when perturbed individually, supporting the hypothesis that the observed double-perturbation phenotypes were the result of genetic interaction. The trade-off inherent in this approach to predicting genetic interactions is that although the analysis can be performed on a genome-wide scale, the nature of each interaction (e. g. additivity, suppression) cannot be predicted. Nevertheless, the approach enables the computational prioritization of potential interaction pairs and greatly reduces the space of likely genetic interactions.

Targeted strategies of data integration have also been used to provide biological context to observed genetic interactions. Kelley and Ideker [45] demonstrated a use for protein-protein and protein-DNA binding networks in understanding the synthetic lethal interaction network in yeast. The study evaluated three standard pathway models for SSL interaction [30,81]: the *between-pathway model*, in which SSL interaction partners are situated in parallel pathways with redundant functions; the *within-pathway model*, in which SSL gene pairs occur for protein subunits of a single complex or pathway that does not lose function until multiple components are removed; and *indirect effects*, in which many pathways are affected and the SSL deletions cannot be associated with specific molecular mechanisms. Based on maximum likelihood, the authors estimated that 40% of yeast SSL interactions could be explained by either the within-pathway or between-pathway models. These models suggest specific patterns of genetic interaction. For the between-pathway model, near-bipartite SSL subnetworks suggest that the missing bipartite links represent undetected interactions. Alternatively, the within-pathway model suggests that genes with many common interaction partners should themselves interact. These predictions were cross-validated by removing one-fifth of the interactions from the data set and then tested to see how well the withheld interactions could be predicted. Both models showed enhanced prediction accuracy relative to randomly predicted interactions, although the between-pathway model was a significantly better predictor, a finding consistent with those suggested by a decision tree approach [88] and an analysis of dSLAM modularity [57,90]. In the worm, integration of protein data also reinforces the compensatory pathway interpretation for SSL interactions [49]. A concurrent study by Zhang et al. [94] integrated molecular and genetic interaction data using the concept of network motifs [53] to guide biological interpretation [92]. In this work, a global network containing protein-protein binding, protein-DNA binding, gene co-expression, sequence homology, and SSL relationships was systematically searched for all possible three-node motifs. A motif in which two proteins linked by protein-protein interaction both have SSL interactions with a third protein was detected in the network ten times more often than random expectation. This was interpreted as a sign of higher-order organization: two proteins that bind in the same complex or pathway share genetic interactions with other proteins in a separate complex or pathway. This interpretation was supported by a search for a four-protein motif in which two parallel pairs of co-binding proteins are linked by all four possible SSL interactions. This motif was found over two hundred times more than random expectation, further supporting the interpretation that SSL interactions are often hallmarks of parallel or compensatory pathways. The goals of integrative analysis are the systematic interpretation of SSL interactions in terms of molecular interactions and the association of specific pathways with observed genetic interactions.

For networks that include multiple modes of genetic interactions, such as suppression, additivity, and genetic synthesis, finding biologically meaningful, non-random network topologies in an integrated network is hindered by the enormous number of possible motifs and the large datasets required for adequate statistics. One strategy for interpreting these rich genetic interaction networks is a quantitative generalization of the classical genetic interaction approach of observing how genetic perturbations interact to affect phenotypes [4]. This method has historically been used to infer how genes influence one another and thereby affect a downstream phenotype, revealing functional relationships such as activation, repression, and pathway order. Using filamentous growth in yeast as a model system (a precursor to invasive growth due to nitrogen starvation), a data decomposition technique was developed to separate genetically "direct" (not necessarily molecularly direct) effects of regulator genes from the genetically "indirect" effects that involve genetic interactions between regulator genes (Fig. 5) [12]. Molecular interaction data were then integrated with the decomposition results to construct regulatory network models of information flow through paths of genes (and gene products) linked by physical interactions. These networks represented quantitative hypotheses of influence from the causal perturbation, through a series of molecular interactions, to the expression of affected genes. Since additional genes were implicated in specific influence-mediator roles in the network, quantitative predictions could be made for gene expression effects arising from their deletion and combinations of deletions. This was done by removing the quantitative influence corresponding to the molecular pathways involving the deleted gene(s). A second round of gene expression data was collected to test a set of predictions and the results demonstrated genome-wide success. Next, by using additional decomposition methods to link gene expression patterns with the filamentous growth phenotype, a network model of whole-cell phenotype regulation was constructed. Predictions for additional perturbations were formulated and experimentally tested. The model successfully implicated new regulators of the phenotype and was used to accurately predict phenotypes for novel combinatorial deletions. While this initial work used laboratory-engineered mutant strains, the techniques can be readily extended to dissect and map complex genetic systems in which a handful of trait genes have been identified. With the increasing availability of human interaction



Systems Genetics and Complex Traits, Figure 5
Influence network of five transcription factors that are trait genes for the filamentous growth phenotype in yeast. *Green* and *red* edges map positive and negative influences, respectively, between genes and from genes to the filamentation phenotype. Edge intensity denotes influence strength. (Adapted from [12])

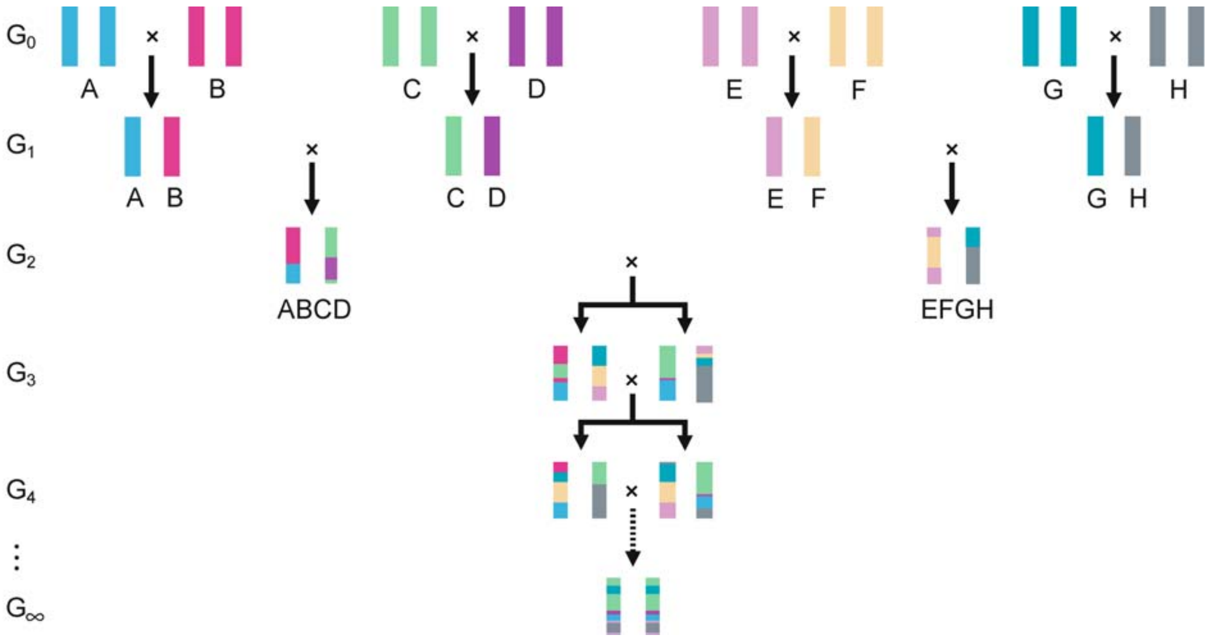
data [73] and further modeling developments to address allelic variation in outbred populations, similar quantitative and integrative techniques may ultimately be applied to disease-related models.

Future Directions

Although genetic complexity is central to all life, systems biology approaches to complex genetics have to this point focused on simpler model organisms. The genetic manipulability, rapid reproduction, and genome-wide interaction databases for yeast, worm, and model prokaryotes have enabled efficient development of experimental and analytical methods. However, the possibility of understanding human disease through the study of complex genetics has catalyzed substantial advances in genetic association studies in mammalian models and human populations [54,55]. The current status of systems approaches to human heath are discussed elsewhere in this encyclopedia (see ► Systems Biology, Introduction to and ► Systems Biology of Human Immunity and Disease), and here we will limit discussion to research opportunities in model systems that are candidates for the systems approaches discussed in this chapter. Future developments can be roughly split into developments in two categories: laboratory resources and computational methods.

Recent programs in rodent breeding have engineered controlled yet genetically diverse populations. These experiments conceptually resemble the analysis of recombinant meiotic progeny generated by crosses in yeast (Sect. "Measuring Genetic Interactions Genome-Wide"). In mouse, two inbred, homozygous parent strains are crossed and two of the resulting F_1 offspring are mated (Fig. 6). The resulting F_2 progeny are then segregated into breeding pairs, and each pair begins a line of repeated sibling inbreeding that continues until a homozygous inbred strain is obtained. Unlike yeast, in which homozygous diploids can be obtained in a matter of weeks, producing recombinant inbred (RI) lines in mice typically requires 20 generations, or 7–8 years, to produce the desired RI lines. Nevertheless, RI genetics have been integrated with gene expression and phenotype assays in mice [10,15,51] and rats [37]. While each of these studies was able to associate individual loci with gene expression patterns, implicating new genes in metabolic, nervous system, and stem cell function, the relatively small numbers of RI strains (approximately 30 in each study) was insufficient for systematically inferring polygenic effects. Inferring such effects would require additional model parameters to encompass the quantitative interaction between each gene pair, leading to overfitting for small data sets. Nevertheless, these works provide prototypes for further studies using RI strains.

While constructing RI strains is expensive and laborious, they have significant advantages over simpler intercross strains in the study of complex genetics. Multiple mice from each RI line will allow biological replication and the observation of isogenic strains under a variety of environmental conditions. Strains can be assayed over long time scales and multiple generations. A collection of shared RI strains would be especially valuable as a community asset both in terms of labor saved, since the strains need only be constructed and genotyped once, and data sharing, since the results from multiple researchers can be integrated in novel analysis. With this in mind, the Complex Trait Consortium is pursuing the Collaborative Cross project [16], which aims to create approximately 1000 RI strains derived from eight inbred founder strains. This set of RI strains will contain a degree of genetic diversity suf-



Systems Genetics and Complex Traits, Figure 6

The collaborative cross strategy for breeding recombinant inbred (RI) strains. The starting parents of the cross (Generation 0, G_0) consists of eight inbred laboratory mouse strains. The goal of the breeding strategy is to create 1000 new RI mouse lines that contain different combinations of alleles from the original parents. A theoretical example of the generation of a single chromosome for a single RI strain is shown. In the first generation of the cross (G_1) progeny receive an unaltered copy of a chromosome from each parent. In all subsequent generations chromosomal material is exchanged between chromosomes generating chromosomes that contain genomic regions from a variety of parents. The lines are ultimately converted to recombinant inbred (RI) strains, in which an individual contains identical copies of these hybrid chromosomes (G_{infinity}) by multiple generations of inbreeding, in this case mating brother-sister pairs. (Reprinted with permission from The Complex Trait Consortium [16])

ficient to uncover gene-gene, gene-drug, and gene-environment interactions. Furthermore, the resulting 1000 inbred mouse lines could be crossed to produce hundreds of thousands of intercross (RIX) strains. Inbred homozygosity ensures that these genotypes will be reproducible and the genetic differences between RI strains will produce low coefficients of inbreeding. Because the RIX strains are not themselves inbred, they will more closely resemble the human population in terms of heterozygosity, genetic mixing, and hybrid vigor. When combined with the proliferation of gene annotation and molecular interaction data, resources like the mice derived from the Collaborative Cross will allow systems approaches to mammalian genetics and provide a platform for adaption to human disease.

In addition to improvements in laboratory resources, a number of analytical advances are required to better understand large-scale genetic interaction networks. These include further developments in data integration, quantitative prediction, data analysis, ontologies, and frameworks for classifying complex genetics.

Strategies of data integration have taken diverse forms. These range from all-inclusive, nearly universal strategies of data pooling [88,95], to integrating only a handful of physical interactions implicated by phenotype modeling [13,45], to modeling constrained by incomplete data resources [52]. The accuracy and reliability of data varies greatly both across data types (e. g. protein-protein versus SSL interactions) and within data types (e. g. protein-protein data from yeast two-hybrid versus co-immunoprecipitation assays). Some methods have used machine learning techniques to weigh data types based on predictive utility [38,94]. Through the repeated application of such methods to multiple data sets and biological systems a consistent picture of data reliability might emerge, providing valuable information for modeling and interpretation of complex genetic systems. Finally, in many cases the integration of published data appears to be performed at a late stage in the work as researchers seek more complete explanations for their results. While this is to some extent an inevitable result of large-scale screens or discovery-driven science, it would be advantageous for investigators to consider from the outset the types of data that may be most complementary and form a coherent plan of data integration [24]. The use of successful mathematical models and biological simulations at such an early stage of research would aid in this integrative research design.

Genetic interaction analysis is often showcased for its predictive modeling potential. Since genetic interactions inherently reveal functional information, under-characterized genes can be assigned to biological pathways [23,71] or functions [80], and these implicit hypotheses about gene function can be experimentally tested. However, physical interaction networks organized into pathways of information flow can also explicitly predict the consequences of additional perturbation [12,66]. The predictions are often quantitative and represent a potentially powerful tool for predictive medicine when applied to the development of therapeutic interventions. Expanded use of computational modeling has the potential to generate more accurate and sophisticated predictions of how systems respond to genetic or environmental perturbations. Established modeling strategies include kinetic modeling, which has been used to understand complex cell circuitry [14,60] and to elucidate subtle stochastic effects [2], and Boolean network approaches, which have been used to model sophisticated systems such as host-parasite interactions [77] and uncover fundamental properties of biological dynamics [68]. Combined with time-series assays of responses to environmental stimuli in genetically diverse organisms, these quantitative modeling techniques can be a powerful tool in understanding the dynamics of polygenic traits.

In the analysis of genetic interaction data, the field of systems genetics now is primarily focused on synthetic lethal interactions, to the extent that many authors use the terms genetic interaction and synthetic lethality synonymously. However, synthetic lethal interactions may be less relevant in the complex genetics of natural populations than in laboratory-based screens. Their broad application may also be limited by reliance on technology specific to model organisms and networks composed of binary interactions. For instance, although synthetic screens can reveal compensatory modules and classify genes into functional pathways, the scale-free architecture of these networks often features many pathways connected to a common hub [49]. Since newly discovered synthetic interactions are most likely to include a hub and a non-hub gene, many pathways will be connected to the hub, and thus be equally likely candidates. It may therefore be difficult to place the latter gene in the correct pathway without additional information. Alternatively, considering a broader spectrum of genetic interactions can both resolve pathway membership, via mutually informative partners, and provide information about pathway structure, e. g. the placement of a factor in a signaling or transcriptional cascade. A network encoding a more general set of genetic interaction rules can thus enable the modeling of specific mechanisms in detail rather than congruence methods based on guilt-by-association. Examples of such rules are directional, targeted interactions like epistasis and suppression (Fig. 4) [23,71] or quantifiable influences describing one gene's influence on another (Fig. 5) [12].

To this point, the classification and interpretation of genetic interactions has been inconsistent due to the paucity of quantitative phenotypes that can be widely screened paired with a limited knowledge of the nature and prevalence of genetic interactions. As the field of complex genetics evolves, so too will the meaning of genetic interaction. To date, genetic interaction studies have primarily focused on pair-wise interactions that are interpreted on a network level. This is often a result of exploring complex genetics with engineered strains, in which single and double genetic perturbations are systematically performed. The combinatorial expansion makes systematic construction of higher-order perturbations (such as triple-mutants) intractable. Indeed, a comprehensive set of double perturbations is daunting, and would be made even more complicated when one considers the corresponding increase in the number and richness of possible phenotypes. In contrast, populations derived from genetically diverse strains, such as yeast and mouse crosses, will sample all genetic combinations of arbitrarily high order. It is possible that network analysis based on pair-wise interactions will be inadequate for experiments involving these strains. Modules or densely interacting "gene complexes" may better describe the space of polygenic effects. With these issues being compounded by the limited capacity to assay many of the phenotypes most relevant to disease, it is currently unknown which approaches will yield the most information about the underlying biology. It is likely that the optimal approach will vary by organism and experiment.

Mature network analyses developed in a broad and systematic framework could be readily expanded beyond genotype-phenotype modeling. To model effects of environment and drugs, additional factors that affect phenotype might be integrated. Potential factors include metabolites, exogenous molecules such as drugs or environmental compounds, and external conditions such as temperature or cell density. A model encompassing multiple factors could quantify how each factor interacts or influences each other factor, in order to produce a map of how effectors interact to generate a phenotype or outcome. Predictive network models focused on specific phenotypes would allow the development of highly specific drugs and aid in the prediction of side effects by observing how a particular perturbation affects connected systems *in silico*. These models could also aid in the identification of allelic combinations that cause genetic disease, the prediction of desired drug-allele interactions for personalized medicine, and the development of multi-drug treatments. Creating such models will require substantial advances in integration of mathematical modeling and genetics, with the focus modest-sized systems rather than genome-scale screens.

In addition to advancing biomedicine, predictive models could also inform evolutionary biology and bioengineering. Rare mutations that confer fitness advantages could be identified by systematically predicting the fitness consequences of perturbations. Perturbations predicted to have effects on multiple phenotypes may uncover evolutionary trade-offs, in which the fitness balance might vary as the environment changes. Fitness assays, such as the E-MAP method in yeast [64], could be used to experimentally test these predictions in model organisms. These effects could in turn be used as input for population modeling and predict evolutionary trends suitable for laboratory testing with fast-replicating yeast or bacteria. As a more practical application, the ability to predict the behavior of re-engineered and synthetic biological networks is a key ingredient in the design and development of novel bioengineered systems. Models that predict complex phenotypes quantitatively and precisely will be a necessary component of elaborate modeling schemes that link genotype, phenotype, and ecosystem.

High throughput data acquisition and the development of computational approaches have enabled the field of systems genetics to progress rapidly. Far-sighted experimental planning and continued technological advances can be expected to fuel continued progress. Current studies, primarily focused on yeast as a model system, are laying the groundwork for network analysis and modeling methods that may be adapted to higher organisms. A target for early adaption may lie in the field of metazoan development to study complex genetic networks involved in multicellular biology, possibly expanding the use of *C. elegans* as a model due to its ease of manipulation by RNAi technology [43,48]. Knowledge accrued from these studies coupled with advances in statistical genetics may form a solid basis for applications in more genetically diverse populations, including the RI mouse strains generated by the Collaborative Cross and their RIX intercross counterparts. The result will be new tools for complex human genetics. Meanwhile, the mapping of the human HapMap, The Cancer Genome Atlas, and human interaction networks will begin to provide data necessary to make use of these tools in personalized and predictive medicine.

Acknowledgments

G.W.C. is supported by grants K25GM079404 and P50GM076547 from NIH and FIBR EF-0527023 from NSF. A.M.D. is supported by a Genome Scholar and Faculty Transition award from the NIH/NHGRI.

Bibliography

1. International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–320
2. Acar M, Becskei A, van Oudenaarden A (2005) Enhancement of cellular memory by reducing stochastic transitions. Nature 435:228–32
3. Altmuller J, Palmer LJ, Fischer G, Scherb H, Wjst M (2001) Genomewide scans of complex human diseases: true linkage is hard to find. Am J Hum Genet 69:936–50
4. Avery L, Wasserman S (1992) Ordering gene function: the interpretation of epistasis in regulatory hierarchies. Trends Genet 8:312–6
5. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, Gifford DK (2003) Computational discovery of gene modules and regulatory networks. Nat Biotechnol 21:1337–42
6. Boone C, Bussey H, Andrews BJ (2007) Exploring genetic interactions and networks with yeast. Nat Rev Genet 8:437–49
7. Brem RB, Kruglyak L (2005) The landscape of genetic complexity across 5700 gene expression traits in yeast. Proc Natl Acad Sci USA 102:1572–7
8. Brem RB, Storey JD, Whittle J, Kruglyak L (2005) Genetic interactions between polymorphisms that affect gene expression in yeast. Nature 436:701–3
9. Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. Science 296:752–5
10. Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT, Wiltshire T, Su AI, Vellenga E, Wang J, Manly KF, Lu L, Chesler EJ, Alberts R, Jansen RC, Williams RW, Cooke MP, de Haan G (2005) Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. Nat Genet 37:225–32
11. Carlborg O, Haley CS (2004) Epistasis: too often neglected in complex trait studies? Nat Rev Genet 5:618–25
12. Carter GW, Prinz S, Neou C, Shelby JP, Marzolf B, Thorsson V, Galitski T (2007) Prediction of phenotype and gene expression for combinations of mutations. Mol Syst Biol 3:96
13. Carter GW, Rupp S, Fink GR, Galitski T (2006) Disentangling information flow in the Ras-cAMP signaling network. Genome Res 16:520–6
14. Chen KC, Calzone L, Csikasz-Nagy A, Cross FR, Novak B, Tyson JJ (2004) Integrative analysis of cell cycle control in budding yeast. Mol Biol Cell 15:3841–62
15. Chesler EJ, Lu L, Shou S, Qu Y, Gu J, Wang J, Hsu HC, Mountz JD, Baldwin NE, Langston MA, Threadgill DW, Manly KF, Williams RW (2005) Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. Nat Genet 37:233–42

16. Churchill GA, Airey DC, Allayee H, Angel JM, Attie AD, Beatty J, Beavis WD, Belknap JK, Bennett B, Berrettini W, Bleich A, Bogue M, Broman KW, Buck KJ, Buckler E, Burmeister M, Chesler EJ, Cheverud JM, Clapcote S, Cook MN, Cox RD, Crabbe JC, Crusio WE, Darvasi A, Deschepper CF, Doerge RW, Farber CR, Forejt J, Gaile D, Garlow SJ, Geiger H, Gershenfeld H, Gordon T, Gu J, Gu W, de Haan G, Hayes NL, Heller C, Himmelbauer H, Hitzemann R, Hunter K, Hsu HC, Iraqi FA, Ivandic B, Jacob HJ, Jansen RC, Jepsen KJ, Johnson DK, Johnson TE, Kempermann G, Kendziorski C, Kotb M, Kooy RF, Llamas B, Lammert F, Lassalle JM, Lowenstein PR, Lu L, Lusis A, Manly KF, Marcucio R, Matthews D, Medrano JF, Miller DR, Mittleman G, Mock BA, Mogil JS, Montagutelli X, Morahan G, Morris DG, Mott R, Nadeau JH, Nagase H, Nowakowski RS, O'Hara BF, Osadchuk AV, Page GP, Paigen B, Paigen K, Palmer AA, Pan HJ, Peltonen-Palotie L, Peirce J, Pomp D, Pravenec M, Prows DR, Qi Z, Reeves RH, Roder J, Rosen GD, Schadt EE, Schalkwyk LC, Seltzer Z, Shimomura K, Shou S, Sillanpaa MJ, Siracusa LD, Snoeck HW, Spearow JL, Svenson K, Tarantino LM, Threadgill D, Toth LA, Valdar W, de Villena FP, Warden C, Whatley S, Williams RW, Wiltshire T, Yi N, Zhang D, Zhang M, Zou F (2004) The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nat Genet 36:1133–7
17. Collins FS, Barker AD (2007) Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies. Sci Am 296:50–7
18. Collins SR, Miller KM, Maas NL, Roguev A, Fillingham J, Chu CS, Schuldiner M, Gebbia M, Recht J, Shales M, Ding H, Xu H, Han J, Ingvarsdottir K, Cheng B, Andrews B, Boone C, Berger SL, Hieter P, Zhang Z, Brown GW, Ingles CJ, Emili A, Allis CD, Toczyski DP, Weissman JS, Greenblatt JF, Krogan NJ (2007) Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. Nature 446:806–10
19. Collins SR, Schuldiner M, Krogan NJ, Weissman JS (2006) A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. Genome Biol 7:R63
20. Davierwala AP, Haynes J, Li Z, Brost RL, Robinson MD, Yu L, Mnaimneh S, Ding H, Zhu H, Chen Y, Cheng X, Brown GW, Boone C, Andrews BJ, Hughes TR (2005) The synthetic genetic interaction spectrum of essential genes. Nat Genet 37:1147–52
21. Dobzhansky T (1946) Genetics of Natural Populations. Xiii. Recombination and Variability in Populations of Drosophila Pseudoobscura. Genetics 31:269–90
22. Donehower LA, Harvey M, Slagle BL, McArthur MJ, Montgomery CA Jr, Butel JS, Bradley A (1992) Mice deficient for p53 are developmentally normal but susceptible to spontaneous tumours. Nature 356:215–21
23. Drees BL, Thorsson V, Carter GW, Rives AW, Raymond MZ, Avila-Campillo I, Shannon P, Galitski T (2005) Derivation of genetic interaction networks from quantitative phenotype data. Genome Biol 6:R38
24. Facciotti MT, Bonneau R, Hood L, Baliga NS (2004) Systems Biology Experimental Design – Considerations for Building Predictive Gene Regulatory Network Models for Prokaryotic Systems. Current Genomics 5:527–544
25. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol 3:121
26. Forster J, Famili I, Fu P, Palsson BO, Nielsen J (2003) Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Genome Res 13:244–53
27. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, Arkin AP, Astromoff A, El-Bakkoury M, Bangham R, Benito R, Brachat S, Campanaro S, Curtiss M, Davis K, Deutschbauer A, Entian KD, Flaherty P, Foury F, Garfinkel DJ, Gerstein M, Gotte D, Guldener U, Hegemann JH, Hempel S, Herman Z, Jaramillo DF, Kelly DE, Kelly SL, Kotter P, LaBonte D, Lamb DC, Lan N, Liang H, Liao H, Liu L, Luo C, Lussier M, Mao R, Menard P, Ooi SL, Revuelta JL, Roberts CJ, Rose M, Ross-Macdonald P, Scherens B, Schimmack G, Shafer B, Shoemaker DD, Sookhai-Mahadeo S, Storms RK, Strathern JN, Valle G, Voet M, Volckaert G, Wang CY, Ward TR, Wilhelmy J, Winzeler EA, Yang Y, Yen G, Youngman E, Yu K, Bussey H, Boeke JD, Snyder M, Philippsen P, Davis RW, Johnston M (2002) Functional profiling of the Saccharomyces cerevisiae genome. Nature 418:387–91
28. Gimeno CJ, Ljungdahl PO, Styles CA, Fink GR (1992) Unipolar cell divisions in the yeast S. cerevisiae lead to filamentous growth: regulation by starvation and RAS. Cell 68:1077–90
29. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG (1996) Life with 6000 genes. Science 274:546, 563–7
30. Guarente L (1993) Synthetic enhancement in gene interaction: a genetic tool come of age. Trends Genet 9:362–6
31. Gunsalus KC, Ge H, Schetter AJ, Goldberg DS, Han JD, Hao T, Berriz GF, Bertin N, Huang J, Chuang LS, Li N, Mani R, Hyman AA, Sonnichsen B, Echeverri CJ, Roth FP, Vidal M, Piano F (2005) Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis. Nature 436: 861–5
32. Hartman JL, Garvik B, Hartwell L (2001) Principles for the buffering of genetic variation. Science 291:1001–1004
33. Harvey M, McArthur MJ, Montgomery CA Jr, Bradley A, Donehower LA (1993) Genetic background alters the spectrum of tumors that develop in p53-deficient mice. Faseb J 7:938–43
34. Harvey M, Vogel H, Morris D, Bradley A, Bernstein A, Donehower LA (1995) A mutant p53 transgene accelerates tumour development in heterozygous but not nullizygous p53-deficient mice. Nat Genet 9:305–11
35. Haugen AC, Kelley R, Collins JB, Tucker CJ, Deng C, Afshari CA, Brown JM, Ideker T, van Houten B (2004) Integrating phenotypic and expression profiles to map arsenic-response networks. Genome Biol 5:R95
36. Huang LS, Sternberg PW (1995) Genetic dissection of developmental pathways. Methods Cell Biol 48:97–122
37. Hubner N, Wallace CA, Zimdahl H, Petretto E, Schulz H, Maciver F, Mueller M, Hummel O, Monti J, Zidek V, Musilova A, Kren V, Causton H, Game L, Born G, Schmidt S, Muller A, Cook SA, Kurtz TW, Whittaker J, Pravenec M, Aitman TJ (2005) Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. Nat Genet 37:243–53
38. Hwang D, Rust AG, Ramsey S, Smith JJ, Leslie DM, Weston AD, de Atauri P, Aitchison JD, Hood L, Siegel AF, Bolouri H (2005) A data integration methodology for systems biology. Proc Natl Acad Sci USA 102:17296–301
39. Ideker T, Galitski T, Hood L (2001) A new approach to decoding life: systems biology. Annu Rev Genomics Hum Genet 2:343–72

40. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 292:929–34
41. Jansen RC (2003) Studying complex biological systems using multifactorial perturbation. Nat Rev Genet 4:145–51
42. Jansen RC, Nap JP (2001) Genetical genomics: the added value from segregation. Trends Genet 17:388–91
43. Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, Welchman DP, Zipperlen P, Ahringer J (2003) Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature 421:231–7
44. Kastenmayer JP, Ni L, Chu A, Kitchen LE, Au WC, Yang H, Carter CD, Wheeler D, Davis RW, Boeke JD, Snyder MA, Basrai MA (2006) Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae. Genome Res 16:365–73
45. Kelley R, Ideker T (2005) Systematic interpretation of genetic interactions using protein networks. Nat Biotechnol 23:561–6
46. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423:241–54
47. Klose J, Nock C, Herrmann M, Stuhler K, Marcus K, Bluggel M, Krause E, Schalkwyk LC, Rastan S, Brown SD, Bussow K, Himmelbauer H, Lehrach H (2002) Genetic analysis of the mouse brain proteome. Nat Genet 30:385–93
48. Lehner B (2007) Modelling genotype-phenotype relationships and human disease with genetic interaction networks. J Exp Biol 210:1559–66
49. Lehner B, Crombie C, Tischler J, Fortunato A, Fraser AG (2006) Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways. Nat Genet 38:896–903
50. Lehner B, Tischler J, Fraser AG (2006) RNAi screens in Caenorhabditis elegans in a 96-well liquid format and their application to the systematic identification of genetic interactions. Nat Protoc 1:1617–20
51. Li H, Chen H, Bao L, Manly KF, Chesler EJ, Lu L, Wang J, Zhou M, Williams RW, Cui Y (2006) Integrative genetic analysis of transcription modules: towards filling the gap between genetic loci and inherited traits. Hum Mol Genet 15:481–92
52. Maslov S, Sneppen K, Eriksen KA, Yan KK (2004) Upstream plasticity and downstream robustness in evolution of molecular networks. BMC Evol Biol 4:9
53. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: simple building blocks of complex networks. Science 298:824–7
54. Monks SA, Leonardson A, Zhu H, Cundiff P, Pietrusiak P, Edwards S, Phillips JW, Sachs A, Schadt EE (2004) Genetic inheritance of gene expression in human cell lines. Am J Hum Genet 75:1094–1105
55. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG (2004) Genetic analysis of genome-wide variation in human gene expression. Nature 430:743–7
56. Ooi SL, Pan X, Peyser BD, Ye P, Meluh PB, Yuan DS, Irizarry RA, Bader JS, Spencer FA, Boeke JD (2006) Global synthetic-lethality analysis and yeast functional profiling. Trends Genet 22:56–63
57. Pan X, Ye P, Yuan DS, Wang X, Bader JS, Boeke JD (2006) A DNA integrity network in the yeast Saccharomyces cerevisiae. Cell 124:1069–81
58. Pan X, Yuan DS, Xiang D, Wang X, Sookhai-Mahadeo S, Bader JS, Hieter P, Spencer F, Boeke JD (2004) A robust toolkit for functional profiling of the yeast genome. Mol Cell 16:487–96
59. Perlstein EO, Ruderfer DM, Roberts DC, Schreiber SL, Kruglyak L (2007) Genetic basis of individual differences in the response to small-molecule drugs in yeast. Nat Genet 39:496–502
60. Ramsey SA, Smith JJ, Orrell D, Marelli M, Petersen TW, de Atauri P, Bolouri H, Aitchison JD (2006) Dual feedback loops in the GAL regulon suppress cellular heterogeneity in yeast. Nat Genet 38:1082–1087
61. Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hon GC, Myers CL, Parsons A, Friesen H, Oughtred R, Tong A, Stark C, Ho Y, Botstein D, Andrews B, Boone C, Troyanskya OG, Ideker T, Dolinski K, Batada NN, Tyers M (2006) Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. J Biol 5:11
62. Rockman MV, Kruglyak L (2006) Genetics of global gene expression. Nat Rev Genet 7:862–72
63. Schacherer J, Ruderfer DM, Gresham D, Dolinski K, Botstein D, Kruglyak L (2007) Genome-wide analysis of nucleotide-level variation in commonly used Saccharomyces cerevisiae strains. PLoS ONE 2:e322
64. Schuldiner M, Collins SR, Thompson NJ, Denic V, Bhamidipati A, Punna T, Ihmels J, Andrews B, Boone C, Greenblatt JF, Weissman JS, Krogan NJ (2005) Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. Cell 123:507–19
65. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 34:166–76
66. Segre D, Deluna A, Church GM, Kishony R (2005) Modular epistasis in yeast metabolism. Nat Genet 37:77–83
67. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–504
68. Shmulevich I, Kauffman SA, Aldana M (2005) Eukaryotic cells are dynamically ordered or critical but not chaotic. Proc Natl Acad Sci USA 102:13439–44
69. Sinha H, Nicholson BP, Steinmetz LM, McCusker JH (2006) Complex genetic interactions in a quantitative trait locus. PLoS Genet 2:e13
70. Spiro PA, Parkinson JS, Othmer HG (1997) A model of excitation and adaptation in bacterial chemotaxis. Proc Natl Acad Sci USA 94:7263–8
71. St Onge RP, Mani R, Oh J, Proctor M, Fung E, Davis RW, Nislow C, Roth FP, Giaever G (2007) Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. Nat Genet 39:199–206
72. Steinmetz LM, Sinha H, Richards DR, Spiegelman JI, Oefner PJ, McCusker JH, Davis RW (2002) Dissecting the architecture of a quantitative trait locus in yeast. Nature 416:326–30
73. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksoz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122:957–68

74. Storey JD, Akey JM, Kruglyak L (2005) Multiple locus linkage analysis of genomewide expression in yeast. PLoS Biol 3:e267
75. Sturtevant AH (1956) A Highly Specific Complementary Lethal System in Drosophila Melanogaster. Genetics 41:118–23
76. Taylor RJ, Siegel AF, Galitski T (2007) Network motif analysis of a multi-mode genetic-interaction network. Genome Biol 8:R160
77. Thakar J, Pilione M, Kirimanjeswara G, Harvill ET, Albert R (2007) Modeling systems-level regulation of host immune responses. PLoS Comput Biol 3:e109
78. Tischler J, Lehner B, Chen N, Fraser AG (2006) Combinatorial RNA interference in Caenorhabditis elegans reveals that redundancy between gene duplicates can be maintained for more than 80 million years of evolution. Genome Biol 7:R69
79. Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H, Andrews B, Tyers M, Boone C (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science 294:2364–2368
80. Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, Chen Y, Cheng X, Chua G, Friesen H, Goldberg DS, Haynes J, Humphries C, He G, Hussein S, Ke L, Krogan N, Li Z, Levinson JN, Lu H, Menard P, Munyana C, Parsons AB, Ryan O, Tonikian R, Roberts T, Sdicu AM, Shapiro J, Sheikh B, Suter B, Wong SL, Zhang LV, Zhu H, Burd CG, Munro S, Sander C, Rine J, Greenblatt J, Peter M, Bretscher A, Bell G, Roth FP, Brown GW, Andrews B, Bussey H, Boone C (2004) Global mapping of the yeast genetic interaction network. Science 303:808–13
81. Tucker CL, Fields S (2003) Lethal combinations. Nat Genet 35:204–5
82. Van Driessche N, Demsar J, Booth EO, Hill P, Juvan P, Zupan B, Kuspa A, Shaulsky G (2005) Epistasis analysis with global transcriptional phenotypes. Nat Genet 37:471–7
83. Vidal M (2005) Interactome modeling. FEBS Lett 579:1834–8
84. Waddington CH (1942) Canalization of Development and the Inheritance of Acquired Characters. Nature 150:563
85. Wayne ML, McIntyre LM (2002) Combining mapping and arraying: An approach to candidate gene identification. Proc Natl Acad Sci USA 99:14903–6
86. Winston F, Dollard C, Ricupero-Hovasse SL (1995) Construction of a set of convenient Saccharomyces cerevisiae strains that are isogenic to S288C. Yeast 11:53–5
87. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, Chu AM, Connelly C, Davis K, Dietrich F, Dow SW, El Bakkoury M, Foury F, Friend SH, Gentalen E, Giaever G, Hegemann JH, Jones T, Laub M, Liao H, Liebundguth N, Lockhart DJ, Lucau-Danila A, Lussier M, M'Rabet N, Menard P, Mittmann M, Pai C, Rebischung C, Revuelta JL, Riles L, Roberts CJ, Ross-MacDonald P, Scherens B, Snyder M, Sookhai-Mahadeo S, Storms RK, Veronneau S, Voet M, Volckaert G, Ward TR, Wysocki R, Yen GS, Yu K, Zimmermann K, Philippsen P, Johnston M, Davis RW (1999) Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science 285:901–6
88. Wong SL, Zhang LV, Tong AH, Li Z, Goldberg DS, King OD, Lesage G, Vidal M, Andrews B, Bussey H, Boone C, Roth FP (2004) Combining biological networks to predict genetic interactions. Proc Natl Acad Sci USA 101:15682–7
89. Workman CT, Mak HC, McCuine S, Tagne JB, Agarwal M, Ozier O, Begley TJ, Samson LD, Ideker T (2006) A systems approach to mapping DNA damage response pathways. Science 312:1054–1059
90. Ye P, Peyser BD, Pan X, Boeke JD, Spencer FA, Bader JS (2005) Gene function prediction from congruent synthetic lethal interactions in yeast. Mol Syst Biol 1:0026
91. Yeang CH, Mak HC, McCuine S, Workman C, Jaakkola T, Ideker T (2005) Validation and refinement of gene-regulatory pathways on a network of physical interactions. Genome Biol 6:R62
92. Yeger-Lotem E, Sattath S, Kashtan N, Itzkovitz S, Milo R, Pinter RY, Alon U, Margalit H (2004) Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. Proc Natl Acad Sci USA 101:5934–9
93. Yeh P, Tschumi AI, Kishony R (2006) Functional classification of drugs by properties of their pairwise interactions. Nat Genet 38:489–94
94. Zhang LV, King OD, Wong SL, Goldberg DS, Tong AH, Lesage G, Andrews B, Bussey H, Boone C, Roth FP (2005) Motifs, themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network. J Biol 4:6
95. Zhong W, Sternberg PW (2006) Genome-wide prediction of C. elegans genetic interactions. Science 311:1481–4